nature

www.nature.com/nature

Vol 453 | Issue no. 7198 | 19 June 2008

Solutions, not scapegoats

Scientific misconduct may be more prevalent than most researchers would like to admit. The solution needs to be wide-ranging yet nuanced.

Many researchers would like to believe that scientific misconduct is very rare. But news reported in this issue (see page 969), and the survey results reported by Sandra Titus and her colleagues on page 980, challenge that comfortable assumption. Titus's team found that almost 9% of the respondents in their survey, mainly biomedical scientists, had witnessed some form of scientific misconduct in the past three years, and that 37% of those incidents went unreported.

The results suggest a research climate in which scientific misconduct, although uncommon, is certainly not an anomaly. Titus *et al.* outline a number of measures to address this situation, including better protection for whistleblowers, and promotion of a 'zero tolerance' culture in which scientists have just as much responsibility to report others' misconduct as they have for their own behaviour.

However, although these proposals have much to recommend them, they are, at best, a beginning. A more radical change of perspective may be in order — one in which misconduct is no longer viewed as problem that can be solved by identifying and banishing a few unethical individuals. Instead, the problem calls for approaches that are both more nuanced and more far-reaching.

Consider, for example, that not all cases of misconduct are equally egregious, and not all perpetrators deserve to be branded as cheaters for the rest of their careers. There is often room for honest mistakes and differences of opinion. Yes, institutions should develop strong guidelines for what is and is not permissible, but officials should also have the flexibility to compare individual situations to these guidelines, and to develop unique solutions as needed. In some cases — for example, a young researcher who simply yielded to temptation once — a system of warnings might be used to both correct the problem and educate the researcher. Within individual labs, moreover, airing complex matters — such as decisions about when data can be justifiably excluded from analysis, or how images can be ethically adjusted to improve their quality — may reduce the chance that any single investigator's decision will later lead to accusations of misconduct.

Meanwhile, misconduct investigations all too often focus solely on an individual offender, and fail to diagnose the environment that has allowed misconduct to flourish. Instead, institutions should seize the opportunity to learn from the experience, and to address the bigger questions. For example, did the atmosphere in the lab create the pressure to cut corners? Or did the intensity of the tenure chase contribute? One way to address such questions might be through internal departmental discussions, in which everyone is free to admit mistakes, and discuss how to fix the problems instead of apportioning the blame.

"Investigations often fail to diagnose the environment that has allowed misconduct to flourish."

More-formal misconduct investigations may need to be kept private, as a necessary safeguard to protect the falsely accused. Nonetheless, institutions can and should share the lessons they have learned from the process. Officials at an institution may learn, for example, that mentoring needs to be improved, or that their system for reporting misbehaviour is flawed. Unfortunately, some institutions may instead feel pressure to bury or cover-up their findings for fear of negative press. But to do so is to gain a short-term reprieve at the expense of long-term loss: such institutions will only be doomed to repeat past mistakes.

This means turning attention away from scapegoats, and focusing on solutions. ■

Change in the weather

A renewed push for scientific research into weather-modification technologies is long overdue.

In the 1956 movie *The Rainmaker*, Burt Lancaster plays a con man catering to the dreams of spinster Katharine Hepburn. And while both stars triumph in the end — the rain does fall, and she comes out of her shell — the implication remains that rain-making is little more than a scam.

Today's rain-makers struggle with their own credibility issues. They do have well established methods for seeding clouds with silver iodide crystals, which in most cases bolster precipitation by a small but significant amount (see page 970). That's enough to make the effort worthwhile for communities looking to bolster the snowpack on which they rely for water in summer, or to target rainfall over an agricultural area rather than a neighbouring one that is barren.

Yet weather-modification supporters face a perceived negative bias in the scientific community. For instance, a 2003 report from the US National Research Council publicly doubted whether weather-modification techniques work at all, although it did call for more investment in the field. There has yet to be the definitive experiment that settles exactly how well cloud seeding — or other weather-modification techniques, such as diverting fog or suppressing hail — works (or not).

Part of the scepticism is due to the field's chequered history. The field was born in a blaze of enthusiasm in the General Electric Research Laboratory in New York in 1946, when researchers began dreaming of weather modification on a grand scale — showering areas with rain and redirecting lightning strikes. But decades passed with little concrete progress; even the United States' wonderfully named Project Stormfury, which aimed to weaken hurricanes before they reached land, fizzled out in 1983. And basic questions remain unanswered regarding cloud and

atmospheric physics, such as the influence of air-pollution aerosols.

As is the case in so many areas, the issues with weather modification boil down partly to an uneven allocation of resources. Some countries, such as Israel, have bucked the trend; the country's early experiments with cloud seeding identified the many scientific unknowns that remain, and the government has continued to fund ongoing work to understand those factors better. Other countries, such as the United States, have simply given up; the most promising experiment in America is run not by the federal government but by the state of Wyoming, which is spending nearly US$9 million on a five-year series of cloud-seeding experiments evaluated by experts from the National Center for Atmospheric Research. That's the type of targeted and rigorous study that needs to be done in weather modification, but it took Wyoming to do it.

Elsewhere, plenty of money is flooding into the field, but on the wrong methods. There is little doubt that China's massive weather-modification undertaking has huge appeal for its rain-starved farmers. But most of the money goes on the operational costs of running technologies that have yet to be validated by science. China has the resources and the willpower to lead the world in weather-modification research, but has not yet stepped fully into that role. One promising move, however, is its newly established centre for weather-modification research.

If researchers could improve their understanding of weather modification, it might then be possible to tackle some of the larger legal and political issues. What happens, for instance, when one country wrings excess water out of a cloud before it drifts over a similarly parched neighbour? How does one engage cross-border negotiations on atmospheric rain, when terrestrial water (in aqueducts and rivers) itself is so contentious? Who actually owns the weather?

The stakes are high, as weather modification is one of those areas in which science can have an immediate and obvious benefit for society. It's long past time to invest modest funds in the basic understanding of it. Otherwise, the world's rain-makers may find themselves considerably less successful than Burt Lancaster. ■

Supporting the future

... but the European Research Council's success is undermined by practices beyond its control.

For most of the past four centuries, Europe has been one of the world's great crucibles of revolution — the place where artists, scientists, philosophers and industrialists overthrew the medieval order and pioneered a new age of democracy, technology and individual initiative. And yet, thanks to lingering cultures of hierarchy and institutional rigidities, continental Europe today is a surprisingly difficult place to be a young scientist. Witness the way so many of those young minds continue to flock westward, either to Britain, the least hidebound European country for young scientists, or to the even greater opportunities in the United States. The few institutions on the continent that have managed to empower young scientists — a notable example being the European Molecular Biology Laboratory in Heidelberg, Germany — remain the exception rather than the rule.

Thus, in a week when an Irish referendum has plunged the European Union into new paroxysms of constitutional uncertainty, there is reason to celebrate a new organ of the European Union that is taking a notable step in the right direction: the European Research Council (ERC), founded early in 2007. The ERC's founders deserve great credit for their determination to keep it independent of political and economic considerations, so that it can award its grants on the basis of scientific excellence alone. The ERC has now done just that, giving out substantial money — with much more to come — to young scientists who can take those funds to a host institution of their own choice (see page 975). A sign of the council's success is that some countries have set up special funding schemes to support high-rated applicants who, because of the massive oversubscription, failed to get ERC funding. Credit, therefore, should go to Chris Patten, the chancellor of the University of Oxford, who chose the initial council members; to Ernst-Ludwig Winnacker, the ERC secretary-general who is responsible for its relations with the European Commission; and to council president Fotis Kafatos, who is responsible for delivering its goals.

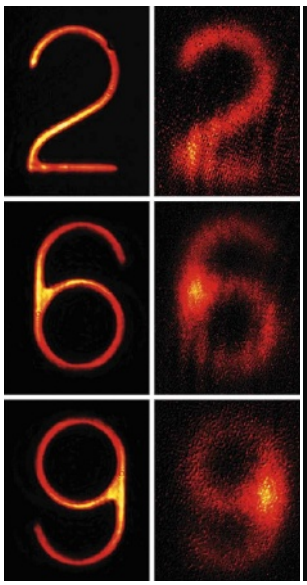
The prestige of the ERC makes it relatively easy to defend; its champions will have little difficulty in obtaining vocal, high-level support if its drive for unmitigated excellence is threatened. Indeed, the avoidance of this threat — so far — has been one of its most notable achievements. It would have been easy for countries new to the European Union, many of which have relatively weak scientific track-records, to resist the evident trend for ERC fundees to take their grants to host institutions in already dominant European countries. But it seems to be widely recognized across Europe that the goal of scientific excellence should trump nationalistic considerations. Élite diasporas can even be viewed as an investment: at least some of the young émigrés will eventually return to their own countries.

"The goal of scientific excellence should trump nationalistic considerations."

None of this, however, is any excuse for complacency. Pressure to hijack the ERC agenda for political ends can be expected from time to time, and must continue to be firmly resisted. And even more important in the meantime is the need to streamline the European Commission's remarkably inflexible bureaucratic arrangements for ERC awards, which threaten to undermine the council's success. Despite doughty championship of the ERC by the research commissioner Janes Potočnik, and the council's quasi-independent status, grantees are being treated like contractors. This means that both the amount and terms of their funding are subject to negotiations that can drag on for months, which makes planning impossible. It also means that talented young scientists can find more ready terms in the United States and vote with their feet — as has already happened in one or two cases.

Thus, a key priority for those in a position to make a difference should be to change either the implementation or the constitution of the ERC, both to preserve and extend what has been achieved so far, and to stop defeat being snatched from the jaws of victory. ■

RESEARCH HIGHLIGHTS



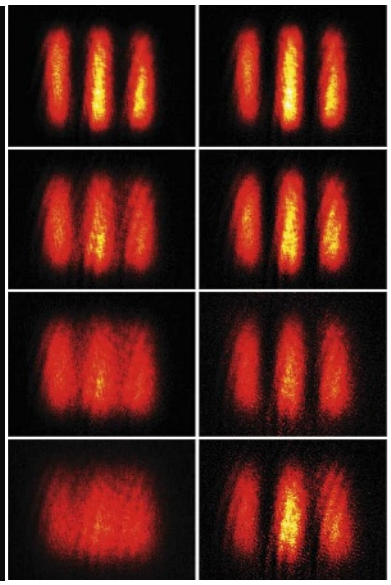
Fuzzy figures

Phys. Rev. Lett. **100,** 223601 (2008)

Capture the complex patterns of photons that make up several numerals in a vapour of rubidium atoms at 52 °C, and those images will degrade as the atoms diffuse (pictured left). But Moshe Shuker of the Technion-Israel Institute of Technology and his colleagues have found a way to store such images and then regenerate the original light beam. The numbers were created by projecting a laser beam through a stencil and exciting the atoms.

Shuker's team stored images comprising sets of three parallel lines for 2, 10, 20 or 30 microseconds (pictured far right and in descending order) using a 'phase shift' technique to counteract the effect of diffusion (shown near right). The technique involves manipulating the phase of the input image, which controls the quantum phases of the atoms. The phases of the atoms that diffuse away from an image's lines are at 180° to one another, and so cancel each other out in the restored image.

Thirty microseconds is a thousand-fold increase over the previous record for delaying an image. The work has potential applications in many fields, including quantum information processing.



M. SHUKER ET AL.

CHEMICAL NANOTECHNOLOGY

Close the gate

J. Am. Chem. Soc. doi:10.1021/ja800266p (2008)

Nanoscale synthetic channels that are opened and closed by a DNA 'switch' have been constructed by a team in China. Such channels could form part of a selective membrane for filtering and purifying water or for mimicking the changeable permeability of biological ion channels.

Yugang Wang of Peking University and his colleagues etched funnel-shaped holes, 5–44 nanometres wide at the narrowest point, into polymer membranes and lined the pores' mouths with single strands of DNA. The DNA in the pore is tightly folded in acidic conditions but unravels into loose chains at pH 8.5. This alters the diameter of the hole and therefore the flux of ions through it.

MOLECULAR BIOLOGY

Sod it

Genes Dev. **22,** 1451–1464 (2008)

Mutations in the *SOD1* gene cause motor neurons to die in amyotrophic lateral sclerosis, also known as Lou Gehrig's disease. Hidenori Ichijo of the University of Tokyo and his co-workers have pinned down why.

The key lies in the system of intracellular membranes called the endoplasmic reticulum (ER). Mutations in *SOD1* seem to affect the system that degrades worn-out pieces of ER, and a surfeit of ER containing misfolded proteins activates a genetic programme that kills the cell.

Ichijo's team found that they could mitigate motor-neuron death and extend the lifespan of *SOD1*-mutant mice by deleting a gene (*ASK1*) that turns on the cell-death programme.

ANIMAL BEHAVIOUR

Token symbolism

PLoS ONE **3,** e2414 (2008)

Apes use and understand symbols but they are not unique in this respect: capuchin monkeys (*Cebus apella*; below) can assign values to tokens that represent different items of food.

Elsa Addessi of the CNR, Italy's national research council, and her colleagues trained five monkeys to associate a particular token — such as a green chip, black plastic tube or a brass hook — with one of three specific types of food. They then gave the monkeys a series of choices, each time between different amounts of two food items or between two types of token.

The value the monkeys assigned to a token was very similar to the value they gave to the food it represented, which suggests that the animals weighed up both real and symbolic options in an equivalent manner.

E. VISALBERGHI

ASTROPHYSICS

Cosmic tiara

Astrophys. J. **680,** 295–311 (2008)

A halo of stars surrounds the Milky Way, but researchers disagree how it got there. One theory proposes that it formed from the same cloud of gas as the galaxy itself; the other says the halo is the remains of several 'dwarf galaxies' that were originally separate from but close to the Milky Way proper. A survey of about three million halo stars weighs heavily in favour of the latter hypothesis.

Eric Bell of the Max Planck Institute for Astronomy in Heidelberg, Germany, and his colleagues compared data from the Sloan Digital Sky Survey with several models. The halo's structure, they say, suggests that it is the remains of several smaller galaxies that were subsumed into the Milky Way after it formed.

ECOLOGY

Dotty diets

Nature Nanotech. doi:10.1038/nnano.2008.110 (2008)

Those who worry about nanotechnology do so partly because of its potential environmental impact. So David Holbrook and a team from the US National Institute of Standards and Technology, in Gaithersburg, Maryland, have tested whether quantum dots (tiny blobs of semiconducting material) accumulate in a simple invertebrate food web.

Over a series of experiments, they put bacteria (*Escherichia coli*), rotifers (*Brachionus calyciflorus*) and ciliates (*Tetrahymena pyriformis*) in flasks with carboxylated and biotinylated quantum dots, which may find a use in computing and solar cells.

The nanomaterials could only stick to clumps of bacterial cells — aggregates too large for ciliates to gobble. However, ciliates took up quantum dots directly from the media, retaining the biotinylated dots for more than twice as long as the carboxylated ones. Rotifers, which eat ciliates, thus consumed quantum dots, but emptied the dots from their guts fast enough to avoid accumulating them.

NEUROSCIENCE

Wide awake

Nature Neurosci. doi:10.1038/nn.2140 (2008)
When it comes to neuronal activity, researchers often assume that what holds for anaesthetized subjects holds for those that are fully awake. This simple inference is misguided, Jason Kerr of the Max Planck Institute for Biological Cybernetics, in Tübingen, Germany, and his colleagues have found.

They recorded how pairs of neurons behave in unmedicated rats and how they behave in the same rats when dosed with ketamine. The neuron pairs that generated the strongest correlations in their discharges before the animals were anaesthetized were not those that were most strongly correlated when the rats were drugged.

This means that care must be exercised when extrapolating measurements of firing patterns across populations of brain cells in the anaesthetized to the wakeful.

MOLECULAR BIOLOGY

Shaping up

Science **320,** 1471–1475 (2008)
How does ubiquitin, a regulatory protein that labels other proteins for destruction, bind to so many different structures? By shuffling between arrangements until it finds the best option, according to Bert de Groot of the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany, and his team.

Forty six of the arrangements were already known from X-ray crystallography of ubiquitin recognition complexes. The researchers followed ubiquitin's structure over pico- to microseconds in various solutions and from many angles, showing that all these conformations are likely to be adopted in living cells.

This work adds to evidence that many confirmations of the same protein often exist in dynamic equilibrium before a binding partner comes along, a model that is at odds with the 'induced fit' hypothesis.

ENVIRONMENTAL MONITORING

Arsenic detectives

Proc. Natl Acad. Sci. USA doi: 10.1073/pnas.0710477105 (2008)
Dissolved arsenic was discovered in the groundwater of the Bengal Basin of Bangladesh and India more than twenty years ago. With deeper wells, safe drinking water might be provided for more than 90% of this region, according to an analysis by Holly Michael and Clifford Voss of the US Geological Survey (USGS) in Reston, Virginia.

The release of arsenic into the basin's groundwater is mainly caused by reduction of iron oxyhydroxides, which tends to take place near the surface. Most wells in the area pump from the contaminated zone, even though the polluted groundwater rarely reaches deeper than 100 metres.

J. HOLMES/PANOS

The USGS model of groundwater flows in the basin suggests that water taken from depths of 150 metres or more will not, in most areas, be tainted by arsenic for a millennium.

MOLECULAR BIOLOGY

Cancer's instigators

Cell **133,** 994–1005 (2008)
Some primary tumours stimulate the spread of cancer by releasing a protein called osteopontin, studies in mice suggest.

Robert Weinberg of the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, and his colleagues implanted tissue from vigorously growing human breast tumours into mice. They then injected tumour cells that normally grow slowly. The fast-growing tumours spurred the enlargement of the 'responder' tumours via osteopontin, which has been previously linked to poor prognosis in several human cancers. Blocking osteopontin's action may yield useful cancer treatments.

JOURNAL CLUB

John P. Quinn
Queen's University, Belfast, Northern Ireland

A microbiologist learns that all marine creatures must suffer for the greed of a few.

Phosphate is an essential nutrient for all forms of life. Demand for it tends to outstrip supply to such an extent that it limits the overall productivity of many ecosystems, including vast tracts of the seas. I study the curious strategies by which creatures obtain sufficient phosphate for life as they know it.

Some microorganisms, for instance, keep a phosphate store for when times are hard. They scavenge for the nutrient in their surroundings with high-affinity uptake systems and then produce polyphosphate, an insoluble polymer that packs hundreds of phosphate subunits into a single strand. Strands of polyphosphate then form intracellular granules that can be broken down by cellular enzymes when they are needed.

This kind of 'luxury' uptake was recently the focus of a study by Ellery Ingall of the Georgia Institute of Technology in Atlanta and his colleagues. Diatoms — unicellular, silica-walled algae — accumulate phosphate during summer blooms to levels far beyond their immediate needs. Indeed, polyphosphate produced by plankton accounted for 7–11% of the total phosphate in the surface waters of Effingham Inlet, a fjord on Vancouver Island, Canada (J. Diaz *et al. Science* **320,** 652–655; 2008).

This self-indulgent behaviour seems to have far-reaching consequences. Decaying plankton eventually sink to the ocean floor, where they spill unused polyphosphate onto the sediment surface. Notably, Ingall and his team found that soluble phosphate was not released at this point. Instead, polyphosphate molecules seeded the precipitation of minerals called apatites, a process that took only a few years. So diatom greed may ultimately lower the ceiling on marine productivity by locking away the oceans' most hard-to-come-by nutrient. That is important as well as curious.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

Japan ramps up patent effort to keep iPS lead

Japan's Shinya Yamanaka pioneered the field of induced pluripotent stem cells.

T. KITAMURA/AFP/GETTY IMAGES

As the battle to create therapeutic stem-cell lines intensifies, Japan is waking up to the fact that the United States could steal a march on it by being the first to commercialize induced pluripotent stem (iPS) cell technology.

Shinya Yamanaka and his colleagues at Kyoto University pioneered the creation of iPS cells, and the technology is seen as something of a national industry — albeit one in its extreme infancy. Like embryonic stem (ES) cells, human iPS cells have the potential to develop into any of the body's cell types, and are expected to have tremendous value in drug screening and for therapeutic purposes. They are easier to produce than ES cells and are not associated with the same controversial source — iPS cells can be derived from adult cells rather than embryonic cells.

On the same day in November 2007 that Yamanaka reported his human iPS cells[1], James Thomson's team at the University of Wisconsin-Madison separately published similar results[2]. The details of any patents applied for by either party are not known — in Japan, as in Europe, a patent is awarded to the researchers who file first; in the United States, the patent goes to the group that can show it invented the technology first (see 'Broader coverage').

Kyoto University stalled over developing a strategy to protect its patents because of a lack of legal expertise on involvement with industry. This has caused much anxiety in the Japanese media, with pundits fretting over what they see as imminent US ascendancy in the field. The *Nikkei Keizai Shimbun* newspaper, for example, notes that presentations on iPS cells by non-Japanese groups had arrived "one after the other" at last week's meeting of the International Society for Stem Cell Research in Philadelphia, adding that "Japan should be leading in iPS technology, but things have taken a turn for the worse".

This might be about to change, though, with the launch of 'iPS Academia Japan', a company set up to manage Kyoto University's iPS patents. The company, which is due to start up within a month, will be backed by around ¥1.2 billion (US$11 million) from a fund created jointly in May by Daiwa Securities Group, the Sumitomo Mitsui bank and NIF SMBC Ventures, a private Japanese equity company. A Daiwa Securities representative says that no return is expected, "at least not in the short term. It is a form of corporate social responsibility."

A central purpose of iPS Academia Japan is "to prevent some group or company from monopolizing iPS technology", says Hiroshi Matsumoto, Kyoto University's executive vice-president, who has been heading the dealings between the university and the investors.

It may already be too late. Kazuhiro Sakurada, who led the research arm of drug company Bayer Yakuhin in Kobe, is now chief scientific officer of iZumi Bio in San Francisco, California, a company set up to commercialize iPS cells. In April, reports claimed that Sakurada had created his own iPS cells in April 2007 while at Bayer Yakuhin, even though his results were not published until this January[3]. There are unconfirmed reports that Yamanaka did not create his first cells until July 2007, and it is not known who would hold the critical patent. Although Yamanaka has patents from his original iPS work in mice, it is not clear whether these patents will cover human iPS cells. Neither Sakurada nor Yamanaka would comment on the issue.

Little is known about iZumi Bio, and this exacerbates Japanese fears. The company is in "stealth mode", according to a spokesperson at Burson-Marsteller, the public-relations firm representing iZumi in Tokyo. It has a skeletal website (www.izumibio.com) that presents only its mission statement: to use "the power of induced pluripotent stem (iPS) cells to transform drug discovery and regenerative medicine". Yutaka Teranishi, who heads Kyoto University's intellectual-property office, says there is currently no formal relationship between the university and iZumi. But he adds: "We would be ready to license the technology to any partner ready to [develop iPS-cell technology for the benefit of patients]."

Thane Kreiner, chief executive of iZumi Bio, told *Nature* only that the company is funded

Broader coverage

The potential patentability of induced pluripotent stem (iPS) cells got a boost last December. The Wisconsin Alumni Research Foundation (WARF), which controls James Thomson's embryonic stem (ES) cell patents, had tried to extend the coverage of its patents to cover all pluripotent stem cells, including iPS cells. The ES-cell patents have been a thorn in the side of many researchers and consumer activists, who feel they are unduly restrictive.

Last year, the US patent office rejected the three patents in question. But in February and March of this year it reversed the decision. The patents had been challenged because it was claimed the technology was the same as that used to derive mouse ES cells. But in the end the patent office recognized differences in the cells. For example, mouse cells express a surface protein, SSEA-1, not expressed in the human version.

The US patent office rejected WARF's claim that the human ES-cell patents cover all pluripotent stem-cell lines, because some pluripotent stem cells show variations (for example, some human germ cells do express SSEA-1). Such physical characteristics may be used to differentiate cell lines in the future, giving patent opportunities to those working on other routes to pluripotency.

"Shinya Yamanaka's iPS lines (and possibly Thomson's iPS lines, depending on who got there first) may fall outside the original WARF patents," says Christopher Thomas Scott of Stanford University's Program on Stem Cells in Society.

WARF is preparing to push the commercial potential of its own iPS patents, as it did with its ES-cell patents. "We are talking to business and industry to gain input for our plan, which we hope to complete this year," says Janet Kelly, WARF communications director. **D.C.**

by Kleiner Perkins Caufield and Byers, the high-powered venture-capital company that helped create Genentech, and Highland Capital Partners. "iZumi is engaging in discussions with various potential partners," Kreiner says, and would not discuss the company's business model further. But on Monday, iZumi announced "a major research collaboration and licensing agreement to focus on applications for iPS cells" with Gladstone Institutes, based in San Francisco. Yamanaka has a joint position there.

Rumours abound in the normally conservative Japanese press. The magazine *Nikkei Biotechnology & Business* reports that iZumi has been collecting "iPS cell patents from all over the world", and has already licensed ES- and iPS-cell-related patents from Harvard University and the Massachusetts Institute of Technology (MIT) in Cambridge. Representatives of the intellectual-property office at MIT deny the report. Counterparts at Harvard say "we have no news to report" with regard to the patents. Kreiner says the patents in question have not even been issued. "It is clearly too early to discuss," he says.

"iZumi would like to work closely with Japan, and we celebrate Japan's scientific leadership," Kreiner told *Nature*. "Yamanaka and Sakurada as well as Thomson have made significant contributions," he says, acknowledging that it was Yamanaka's work on iPS cells in mice that was the starting point for the whole field. ■

David Cyranoski

1. Takahashi, K. *et al. Cell* **131,** 861–872 (2007).
2. Yu, J. *et al. Science* **318,** 1917–1920 (2007).
3. Masaki, H. *et al. Stem Cell Res.* **1,** 105–115 (2007).

VISIONS OF AMERICA/ALAMY

ALL THE RAGE
Why bumper stickers are linked to aggressive driving.
www.nature.com/news

Institutes in pharma cash probe

Until about three years ago, researchers at Duke University in Durham, North Carolina, declared their financial conflicts of interest by filling in a simple form. "It basically just asked: do you have a relationship with a company that might entail a conflict of interest?" says James Siedow, Duke's vice-provost for research.

The form has since been changed every year, becoming more complex with each new iteration. Now, Duke researchers are asked to estimate how much money they receive from industry sources. If that value is more than US$25,000, they are required to delve into specifics. Although Siedow believes that most researchers filled in their forms accurately, he suspects that some were not forthcoming. "There were folks who didn't check that '$25,000 or more' box, yet we're almost certain they should have," he says.

Pressure has been mounting on universities and hospitals to crack down on conflicts of interest. This month, Senator Charles Grassley (Republican, Iowa) informed Congress that three high-profile psychiatrists at Harvard Medical School may have failed to disclose a total of about $4 million of industry earnings over the course of seven years. Last August, Grassley disclosed that a researcher at the University of Cincinnati in Ohio had under-reported industry earnings from a company that made a drug she had tested in clinical trials. And Grassley has informed the National Institutes of Health (NIH) that he is investigating other cases at more than 20 different institutions.

The NIH has made its stand on the issue clear: it is up to individual universities and hospitals to monitor their researchers, says deputy director of extramural research Norka Ruiz Bravo. The NIH may punish individual investigators by abridging or terminating their grants, but in cases where the misconduct seems to be systemic, sanctions could be levied against an entire institution.

But university administrators say they are unable to verify what their researchers disclose. "It's an honour system," says Robert Alpern, dean of the Yale School of Medicine. "We rely on the faculty to tell us the truth. And to be honest, up until a few months ago, I think we all thought they were telling us the truth."

Grassley has proposed legislation, called the Physician Payments Sunshine Act, that would require manufacturers of drugs and medical devices to disclose how much they pay doctors. The act would present institutions with a way to verify the sums their researchers have declared, and both the American Association of Medical Colleges and PhRMA, a lobbying group for the pharmaceutical industry, have praised the proposed act for increasing transparency. But the current draft of the legislation applies only to physicians and leaves out those researchers who do not practise medicine. Philip Pizzo, dean of Stanford Medical School, says he supports the act but acknowledges that it will come at a price. "This will not be an easy or inexpensive process," says Pizzo, who notes that schools may need to hire additional staff to process the new data.

Meanwhile, Siedow says that Duke administrators plan to interview researchers they suspect of not disclosing their earnings. Grassley's recent findings have bolstered Duke's efforts to improve its reporting system. "We can see the writing on the wall in Congress," says Siedow. "We don't want to find out from a pharmaceutical company that one of our researchers made half a million dollars that we didn't know about." ■

Heidi Ledford

See Editorial, page 957.

R. FRIEDMAN/CORBIS

Three high-profile faculty members at Harvard Medical School face allegations of not disclosing hefty industry earnings.

Q&A

Lab disinfectant harms mouse fertility

Two chemicals widely used in cleaning agents for homes, offices and hospitals cause birth defects and fertility problems in mice whose cages have been in contact with them, according to **Patricia Hunt** at Washington State University in Pullman. The quaternary ammonium compounds ADBAC (*n*-alkyl dimethyl benzyl ammonium chloride) and DDAC (didecyl dimethyl ammonium chloride) were identified after an exhaustive search for what was causing a massive drop-off in mouse fertility after Hunt moved her research animals to Pullman from Case Western Reserve Medical School in Cleveland, Ohio, in 2005. The chemicals were in the disinfectant Virex* used in the facility. It is Hunt's second accidental foray into toxicology. In 2003 she linked a rash of mysterious egg defects in her research animals to bisphenol A, a chemical that began leaching from plastic water bottles after a high-pH floor detergent was mistakenly used to clean them. Hunt, who studies mammalian egg development, announced her latest results at the Society for the Study of Reproduction meeting in Kona, Hawaii, last month.

What alerted you to the problem?
After the move we began to experience breeding problems in our mouse colony. Only about 10% of females that were mated in one experiment got pregnant, and of those a large number of late-stage fetuses died. This is very unusual in mice. There were also discrepancies in the developmental ages. Some litters were accelerated, some litters were delayed. And we saw more birth defects in the first few months of our study than we had seen in our previous 13 years at Case.

In our breeding colony, the pups were very small at weaning, which pointed to a lactation problem. And we were losing a lot of mothers during birth. This was probably the same problem of late-stage fetal death; dead fetuses were blocking the exodus of fetuses further up the uterus so the mothers couldn't give birth.

We then superovulated the females, trying to get larger numbers of eggs and embryos. We got a few more embryos, but they seemed to be moving through the reproductive tract too fast. We were picking up eight-cell embryos out of the uterus.

This isn't your first foray into toxicology, so was it easy to spot the culprit?
I guess that I was kind of cocky at the outset. But it took us a year to sort out because we went through all the variables and we just could not figure out what was going on. It was only luck that led us to look at the cages and what might be contaminating them.

How did you identify the cause?
I asked the guy who runs mass spectrometry to tell us which chemicals were present in swabs from our cages. He found that every cage we washed in our cage washer came out with the signature of quaternary ammonium compounds in the Virex disinfectant. These compounds build up in the environment and it was very hard for us to get rid of them. It took months and months for their levels to drop, even after we stopped using quaternary ammonium disinfectants in the facility.

WSU PHOTO SERVICES

What did you use instead?
We went back to the disinfectant that we had been using at Case. It's a chlorine-dioxide-based sanitizer called Clidox.

Didn't other researchers at Washington State have the same experience?
Some of them had problems that were less severe. But they didn't have the same micro-isolator cages — their mice were in conventional cages. And our mice were probably being exposed to higher levels. It was probably a slippery slope for the other investigators: they had seen a slow drop in production, slow enough that most didn't notice. But we were walking out of one environment into another. We realized our productivity was nothing like it had been.

Why haven't you published your data?
We've never been able to run controlled experiments that demonstrate beyond all doubt that this is what's going on, because every time we try to do it we end up re-contaminating the environment in the facility.

We did a side-by-side control. We exposed ten cages and left ten cages, but we found after several months that we had reintroduced the contaminant into our clean cages through the washer. We tried hand-washing the cages outside the facility, but the variables were too hard to control.

What's the response been to your findings?
I've had several people tell me they thought I'd spotted the answer to their problems. There are probably a lot of other investigators whose work has been jeopardized or altered by these types of exposures. Two mouse facility managers at Pullman told me they already knew it impacted on breeding performance, because they had seen it over the years and had removed the chemicals from their facilities.

Are these chemicals widely used?
They've crept into use everywhere. They don't smell, they don't leave a residue. They're valuable in some places, like hospitals, because they kill germs. But we don't necessarily need them in the home.

Do you think they might harm humans?
What concerns me is that they persist for so long in the environment. Given our experience, I am concerned that they might have a deleterious effect on the ovary, uterus and in lactation. This group of compounds acts on the cell membrane, and does a fantastic job of killing everything. But, you know, we're composed of membranes too.

I think the effects we've seen are very significant and potentially important for human health and reproduction, so I'd like to see someone research it. It does not affect my favourite part of the process, which is meiosis. So there's no compelling reason for me to continue investigations. ■

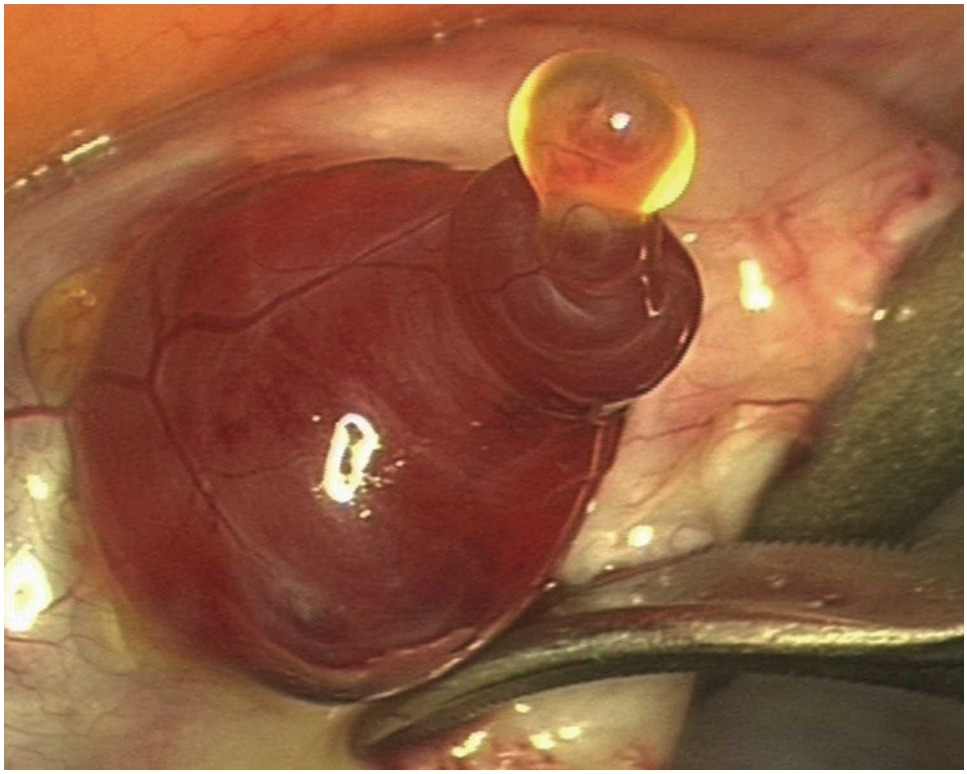
Interview by Brendan Maher

*A spokesperson for Johnson Diversey, which makes Virex, says: "The current abstract is the first we have heard of that attempts to establish a correlation between quaternary ammonium compound disinfection residues and reproductive or developmental effects in laboratory mice. The US Environmental Protection Agency review of these compounds has never indicated any concerns regarding reproductive effects. We will continue to monitor research about the safety profile of quaternary ammonium compounds."

THREE OF A KIND
Trio of 'super-Earths' spotted.
www.nature.com/news/

ESO

SNAPSHOT
Out of the ovary



J. DONNEZ, J.-C. LOUSSE/ELSEVIER

A human egg is caught on camera as it emerges from a woman's ovary. It is one in a series of the clearest ever images of the ovulation process and was accidentally snapped by surgeon Jacques Donnez, of the Catholic University of Louvain in Brussels, as he performed a hysterectomy. The yellowish egg (about the size of this dot ·) is seen exiting a red, fluid-filled follicle on the surface of the 45-year-old woman's ovary.

Ovulation took a full 15 minutes — much slower than the sudden 'explosive' process suggested by some theories. Although human ovulation has never been seen in such detail before, it is a process that occurs at least once a month in fertile women. On release, an oocyte travels along a fallopian tube to the uterus.

These rare images, which appeared in *New Scientist* magazine last week, will be published in the journal *Fertility and Sterility* (J.-C. Lousse and J. Donnez doi:10.1016/j.fertnstert.2007.12.049).

Astronomical wordplay keeps them guessing

Huge Applet, Unsearchable Terrestrials! This anagram could hold the clue to an important extrasolar planet discovery. When astronomer Gregory Laughlin of the University of California at Santa Cruz posted it on his blog last month, readers immediately leapt on the puzzle, and speculation about the hidden message is rife (http://oklo.org/?p=279).

The tradition goes back centuries. Italian astronomer Galileo Galilei embedded his discovery of Saturn and of the phases of Venus in anagrams; Dutch astronomer Christopher Huygens used the same trick to describe his recognition of Saturn's rings (see 'Planetary games'). The device enabled them to stake a claim to a discovery while they slogged through the months of tough observational work needed to confirm the initial idea. Only then would the solution be revealed.

Astronomical historian Owen Gingerich of the Harvard–Smithsonian Center for Astrophysics in Cambridge, Massachusetts, says Laughlin is doing the same thing. "It is just as it was for Galileo: a way of guaranteeing a statement of priority," Gingerich says. "But at the same time, in case he's wrong he doesn't need to decode the anagram."

Laughlin says he's not so much staking a claim as trying to write an interesting blog and draw attention to a golden age of astronomy, when getting scooped by one's peers was a real possibility. "The book of nature is open for anyone to read if they can decode the message," he says.

Gingerich says astronomy is still competitive, especially in the search for planets outside the Solar System. "There have been some serious scraps in this business," he says, adding, "I haven't heard of a biologist putting out an anagram."

Laughlin is keeping mum about the actual discovery while he runs numerical computer models that could confirm it. He's not even sure how noteworthy it will be, given how quickly astronomers are discovering extrasolar planets, which now number close to 300.

Just this month, astronomers announced the discovery of a planet only three times the mass of Earth. There was little reaction in the popular press, nothing like the fuss when the first Jupiter-sized planets were discovered a decade ago. "Now it's kind of old hat to find another one," Gingerich says. ■

Eric Hand

Planetary games

NASA/JPL

Galileo Galilei, ~1610, on Venus (pictured)
Anagram: *Haec immatura a me iam frustra leguntur oy* (These immature ones have already been read in vain by me.)

Solution: *Cynthiae figuras aemulatur Mater Amorum* (The Mother of Loves [Venus] imitates the figures of Cynthia [the Moon].)

Christopher Huygens, 1656, on Saturn
Anagram: aaaaaaaccccdeeeeehiiii iiiillllmmnnnnnnnnnnoooooppqrrstt tttuuuu
Solution: *Annulo cingitur, tenui, plano, nusquam cohaerente, ad eclipticam inclinato* (It is surrounded by a thin flat ring, nowhere touching, and inclined to the ecliptic.)

Gregory Laughlin, 2008, on an extrasolar planetary discovery
Anagram: Huge Applet, Unsearchable Terrestrials!

Solution: to be revealed ...

WORD WATCH

Plutoid
The new name for objects such as poor little Pluto, which was thrown out of the planet club by the International Astronomical Union in 2006. Sidelines thinks that is slightly nicer than the term 'dwarf planet', but also slightly unfair to Eris, the other officially recognized dwarf.

SCORECARD

Irony
Kansas State University's wind erosion lab was blown away by a tornado.

Iron
The rusty metal ranked only sixth in a list of the element names most often mentioned in songs. The top five in the study, which the author admits was more an ode to music and science than a comprehensive analysis, were silver, gold, tin, oxygen and copper.

NUMBER CRUNCH

1 million purpose-made Erector-set pieces are estimated to have been used to build a 'toy' skyscraper at the Rockefeller Center in New York by artist Chris Burden.

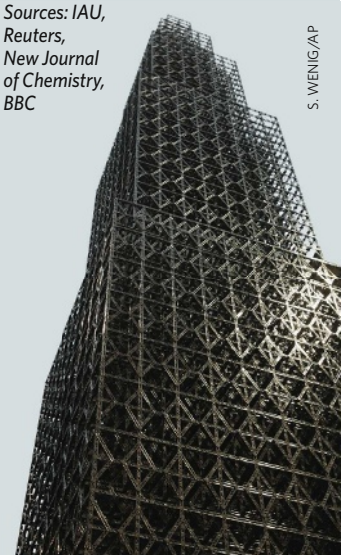
20 metres is the height of the giant building.

7,250 kilograms is its weight.

30 people with screwdrivers were needed to put it together.

1 year was about how long it took them to do it.

Sources: IAU, Reuters, New Journal of Chemistry, BBC



S. WENIG/AP

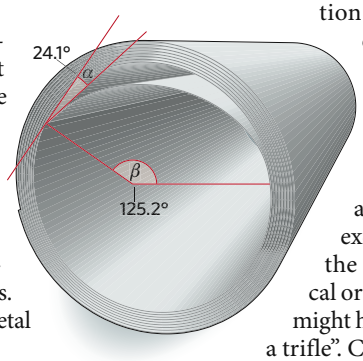
D. WHITEHEAD/CORBIS

Universal law of coiling

Ever noticed that when a piece of paper is rolled into a tube, the innermost part straightens away from the coil before touching down? Try it and see. A team of researchers has investigated this phenomenon and discovered that the precise shape of this rolled-up material is not only surprisingly subtle but also universal.

The angle that the innermost sheet makes with the coiled roll (α in the diagram) is always the same, say Enrique Cerda of the University of Santiago in Chile and his co-workers, about 24.1° — regardless of the thickness of the sheet or the width of the coil[1].

What's more, the angle subtended between this contact point and the place where the sheet first detaches from the coil's inner face (β) is always 125.2°. This universal shape confounds the intuition that stiffer sheets would have a different cross-sectional profile from flimsy ones. Rolled-up carpet, paper or metal will all adopt the same shape.



To prove it, Cerda's team measured the 'touchdown' angle for a thin slab of mica (a sheet-like mineral) and a strip of metal coiled within tubes of various widths. They found that the angles deviated from the predicted 24.1° by no more than about a degree.

"Universal angles have come up before in other situations that involve thin sheets and filaments," says Lakshminarayanan Mahadevan of Harvard University, who was not involved in the work. For example, he and Cerda have previously calculated the universal shapes of flat sheets confined in cylinders by conical deformation, as generated by pushing down on the sheet with a pencil tip[2]. Universal shapes arise, he says, "because of the strong constraints that geometry imposes on the possible deformations".

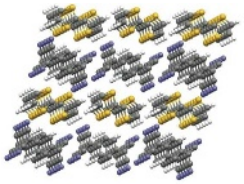
"This type of constraint occurs in other systems as well," Mahadevan explains. "For example, the characteristic size of a drop that breaks off from a stream of fluid always has a size that is comparable to the filament diameter, irrespective of the material of the fluid."

The work is the kind of basic mechanics that one might have expected to have been done already. It is true that the problem is mathematically daunting, involving the calculation of forces and torques that create mechanical equilibrium in a curved, elastic sheet pressing outwards against a confining tube. But Mahadevan says, "the question could have been addressed a long time ago, except for the fact that some of the equations require numerical or graphical solution — which might have slowed things down just a trifle". Cerda says that the phenomenon was well within the reach of eighteenth-century mathematics, but "it seems no one thought to ask".

The researchers say that analogous shapes should exist for coiled sheets in other confining geometries, such as cones or coiled fibres. The coiling of fibres might be relevant to the packing of DNA inside the protein capsules of viruses, and to the mechanisms of biological structures that provide cellular scaffolding. ■

Philip Ball

1. Romero, V., Witten, T. A. & Cerda, E. *Proc. R. Soc. A* doi:10.1098/rspa.2007.0372 (2008).
2. Cerda, E. & Mahadevan, L. *Proc. R. Soc. A* **461,** 671–700 (2005).



ELECTRIC PERFORMANCE
Organic insulators transformed into metal-like conductors.
www.nature.com/news

UK universities in bed with the military

Universities in the United Kingdom may be doing far more research for the military than official estimates acknowledge, according to a report released last week.

Scientists for Global Responsibility (SGR), a Folkestone-based group that campaigns against military spending, says that of 13 universities surveyed, 12 received an average of around £2.4 million (US$4.7 million) each to conduct military and security-related research between 2005 and 2006. Some received as much as £5 million. The figures contrast sharply with SGR's estimate of an average of £400,000 per UK university based on the official 2004 figure of a total of £44 million defence-related research grants across all UK universities. "Our analysis leads us to ask whether government statistics in this area are as reliable as they should be," the study says.

In many cases the money came from both government and commercial sources. Defence firms, including UK-based BAE Systems, Rolls Royce and QinetiQ and US-based Lockheed Martin, make significant contributions to university funding. Most of the money goes to engineering and the physical sciences. In addition, the University of Cambridge, which publishes information on its funding sources, received nearly £1 million from US military organizations. Accurate estimates for US government funding at other UK institutions could not be obtained.

The lack of cooperation by university officials during the study was as disturbing as the actual figures, says Chris Langley, a consultant for SGR, who compiled the report with his colleagues. Vice-chancellors and other senior university personnel refused to speak about the money, and researchers receiving military funding were half as likely to respond to the group's survey compared with other researchers, he says. "We found a huge amount of reluctance." Langley says that he would like to see "safeguards and standards" that would help regulate university money coming from defence sources.

"We do not accept the claim that universities are insufficiently accountable," counters Rick Trainor, president of Universities UK, which represents 132 universities. "Staff and students within universities are expected to adhere to the highest standards of conduct and ethical behaviour in research," he adds.

Langley says that ultimately he believes higher education should spend less on improving the UK's weapons technology and more on research of benefit to society: "We have our priorities wrong," he says.

The Ministry of Defence (MOD) did not comment on the figures, but says it is actively broadening its science and technology supplier base. "MOD is actively seeking to make access to the £500 million a year research budget easier for universities," a spokesperson told *Nature*, adding that there will be further incentives for "academia to engage with the MOD and the established defence industry". ■

Geoff Brumfiel

Beavers are capable of devastating ecosystems (inset).

J. ETCHEVERRY/ALAMY

Tierra del Fuego: the beavers must die

Industrious, shy herbivores they may be, but the beavers of the Tierra del Fuego archipelago on the southern tip of South America are such a menace that scientists are planning the largest eradication project ever attempted.

In the 1940s, 50 North American beavers (*Castor canadensis*) were introduced to the area by the Argentine government to help start a fur industry — their numbers have now swelled to an estimated 100,000. The aquatic rodents, which have thrived in the absence of native predators, have invaded roughly 16 million hectares of unique, indigenous forest, leaving a swath of destruction "that is absolutely stunning — it looks like bulldozers steamed through", according to ecologist Josh Donlan, director of Advanced Conservation Strategies, a non-profit organization based in Driggs, Idaho.

"We will move in on the beavers in a rolling front."

Although North American trees have evolved with beavers and many are able to grow back from their roots, South American trees, such as beeches, simply die when the animals gnaw them down. The dams the beavers make turn stream areas into stagnant bogs, leaving a huge impact on aquatic life, says ecologist Christopher Anderson of the Institute of Ecology and Biodiversity in Santiago, Chile. When these ponds drain out, the muddy areas become meadows that then invite exotic species. "The change in the forested portion of this biome is the largest landscape-level alteration in the Holocene — that is, approximately 10,000 years," Anderson says.

The Argentine and Chilean governments are now reviewing a feasibility study on a total eradication of these beavers, which was undertaken by an international team including Donlan. It would be an eradication over the largest area ever attempted "by an order of magnitude", Donlan says. Beaver-control projects, such as killing traps, are now being ramped up in a bid to test eradication methods.

The dam busters

One priority is preventing the beavers from going north — a few beavers have already been spotted on mainland Chile. To save these southernmost forests, the beavers "must be totally eradicated", says forest engineer Guillermo Martínez Pastur at the Austral Center for Scientific Investigation in Ushuaia, Argentina. But Pastur believes such an eradication "is impossible, or is of extremely high cost" because the area is extraordinarily rugged and remote.

Still, Donlan thinks it is feasible. "We've made huge progress over the past five years in removing invasive mammals from islands," he says, citing the recent eradication of some 140,000 goats from more than 500,000 hectares in the Galapagos Islands. The most likely scenario would be to go in with trappers and dogs using helicopters and boats, and adapting techniques from beaver control in the United States and Canada, Donlan says. "We'll have to move in on the beavers in a rolling front, going from watershed to watershed to remove them, with a massive monitoring programme behind it to make sure they have all been eradicated."

Anderson, who reviewed the feasibility study, also thinks eradication will be possible. "The beavers only live near the water, so you don't have to go over the whole landscape," he explains. "But you have to make sure you eradicate them all — if you have even two beavers, they could repopulate the whole archipelago."

On the opposite side of the globe, a reintroduction programme is scheduled for next year. The European beaver is to be released into Scotland where it has been extinct for 400 years, environment minister Michael Russell announced last month. "Other parts of Europe, with a similar landscape to Scotland, have reintroduced beavers and evidence shows that they can have positive ecological benefits, such as creating and maintaining a habitat hospitable to other species," he told reporters. ■

Charles Choi

C. MONTEATH/HEDGEHOG HOUSE/MINDEN PICTURES/GETTY

Massachusetts finally passes life-sciences bill

The Massachusetts legislature has approved a 10-year, US$1-billion life sciences bill aimed at attracting biotechnology companies to the state. The bill has been championed by Governor Deval Patrick, who signed it into law on 16 June.

The legislation will fund several university research centres, $250 million in tax credits and $250 million in grants, among other projects.

Although legislators overwhelmingly supported the bill, it has generated some controversy. Economists voiced scepticism after Patrick claimed that the funding would generate 250,000 jobs in the state — a figure many considered overly optimistic. While under consideration in the legislature, the bill picked up several controversial additions, including a proposed $50-million science centre at the Massachusetts College of Liberal Arts in North Adams, which does not have a graduate programme in science. That earmark has since been removed.

Anti-AIDS vitamin advertising banned

A South African judge has ruled that clinical trials purporting to test the effectiveness of vitamins against AIDS were illegal, and that vitamins should not be advertised as therapies for AIDS.

The ruling covers 12 defendants, including German doctor Matthias Rath and American doctor David Rasnick, a former advisor to President Thabo Mbeki. The pair had been selling vitamin therapies as AIDS treatments, advertising them as effective and saying that anti-retroviral drugs were toxic.

The lawsuit was brought by the South African advocacy group Treatment Action Campaign and the South African Medical Association, which argued that the defendants had violated national regulations covering medicines and related substances.

During the trial the country's health minister, Manto Tshabalala-Msimang, argued that the therapies should not be covered by the regulations as they were not medicines.

AFP/GETTY

Demonstrators campaigned for Matthias Rath's clinical trials to be judged illegal.

Virtual butterflies

LINNEAN SOCIETY, WWW.LINNEAN.ORG

This *Papilio paris* L., or Paris peacock butterfly, is one of the digitized images being published online from the collections of the Linnean Society of London.

Butterflies and moths — many of them reference or 'type' specimens named by Carl Linnaeus — are being added to celebrate Britain's National Insect Week in late June. The society has already put plant specimens and correspondence online at www.linnean.org.

The Paris peacock is a swallowtail butterfly native to southeast Asia.

Researcher suspended for falsifying data

The Ottawa Health Research Institute last week suspended postdoctoral fellow Kristin Roovers after learning that she had manipulated and falsified data published in several papers.

Roovers was hired by the institute in 2005. But in July 2007, the US Office of Research Integrity concluded that Roovers, while a graduate student and postdoctoral fellow at the University of Pennsylvania in Philadelphia, had manipulated 19 panels of western blot data. She had used Photoshop to copy a set of bands and paste them into other blots representing data from different experiments. The data ultimately appeared in 11 figures in three publications.

Two of the papers (K. Roovers and R. K. Assoian *Mol. Cell. Biol.* 23, 4283–4294; 2003, and K. Roovers *et al. Dev. Cell* 5, 273–284; 2003) have been retracted. A decision on the third (C. F. Welsh *et al. Nature Cell Biol.* 3, 950–957; 2001) is pending.

The Office of Research Integrity barred Roovers from receiving any US government grants for five years.

See Editorial, page 957.

Stem-cell society condemns medical tourism

In a strike against stem-cell tourism, the International Society for Stem Cell Research (ISSCR) last week officially condemned unproven stem-cell treatments that lack appropriate oversight or patient monitoring, or use poorly characterized cells. The society advised clinicians to refuse to participate in nonconforming activities "as a matter of professional ethics".

To help patients make better-informed decisions, an ISSCR task force is drafting guidelines to define when patients should receive stem cells or their derivatives. The draft guidelines were scheduled for release last Thursday, but after disagreements about specificity and tone, the task force decided to announce general principles instead.

The society will not enforce them or evaluate individual clinics for compliance. Instead, the hope is that the guidelines will push countries to adopt and enforce harmonized regulatory standards.

Guidelines cover how cells should be processed and characterized, what preclinical evidence should be collected and what information patients must receive. They also say that stem-cell researchers should consider how their work stands to benefit society as a whole.

Lab-equipment giants set to merge

Two California firms that produce tools vital to the biotechnology industry will merge by the autumn.

Invitrogen of Carlsbad is acquiring Applera's Applied Biosystems Group of Foster City for US$6.7 billion in cash and stock. If approved by shareholders, the combined operations will be called Applied Biosystems, with corporate headquarters in Carlsbad.

Applied Biosystems, with annual revenue of $2.1 billion, is widely known for producing the genetic-sequencing tools used in the Human Genome Project. Invitrogen, with $1.3 billion in annual revenue, is a major supplier of reagents, cells and test kits.

Invitrogen's Gregory Lucier will be chairman of the board and chief executive of the new firm, with Applied Biosystems' Mark Stevenson as president and chief operating officer.

NATURE|Vol 453|19 June 2008

TAMING THE SKY

Is it really possible to stop rain, invoke lightning from the heavens or otherwise manipulate the weather? **Jane Qiu** and **Daniel Cressey** report on the once-scorned notion of weather modification.

China wants everything to be under control at the opening of the Beijing Olympic Games on 8 August — even the weather. The chance of rain that day is 47%, according to the Beijing Meteorological Bureau. The iconic 91,000-seat main stadium, nicknamed the 'bird's nest' because of its interlacing steel beams, has no roof. So Chinese meteorologists will use weather-modification technologies to try to stop rain from spoiling the party.

Beijing's plan for the games is the most conspicuous example of the country's massive weather-modification efforts. Most of the time, the focus is not on keeping things dry, but on making it rain in places that desperately need the water. In ancient times, the Chinese believed that dragons controlled the weather, and elaborate rituals were performed to bring about sufficient rainfall and good harvests. Today, they are turning to technology to change the moods of the sky, part of a national obsession to 'tame nature', as championed by Mao Zedong.

China has one of the largest programmes for weather modification in the world. It spends between 400 million yuan (US$60 million) and 700 million yuan a year on it, and employs 32,000 people to operate 35 specially equipped planes, 7,000 anti-aircraft cannons and 5,000 rocket launchers. Official figures from the China Meteorological Administration say that the country created 250 billion tonnes of rain between 1999 and 2006, an annual production of more than 30 billion tonnes. This is enough to meet the needs of more than 500 million of its 1.3 billion people, but the country aims to generate 50 billion tonnes a year by 2010.

Many researchers, both in and outside China, doubt that sufficient evidence has been accumulated to support this claimed success. "In fact, China is very much behind in this area," says Zhang Hong-fa, an atmospheric scientist at the Cold and Arid Regions Environmental and Engineering Research Institute in Lanzhou. "A false sense of achievement would impede genuine progress."

China also faces long-standing scepticism about weather modification in general. Proponents argue that it's possible not only to produce more rain, but also to get rid of fog, prevent hail and even divert hurricanes from making landfall (see 'Forget the weather forecast', overleaf). Critics say that many of these claims are laughable, and that most of the projects under way are based on little more than faith.

Even supporters are dubious about what China may be able to pull off. "The concern for me is that the Chinese are promising they're going to do something that they really don't have optimum assurances they can deliver," says Bruce Boe, director of meteorology at Weather Modification in Fargo, North Dakota, one of the world's largest weather-modification companies. "This has been something that has plagued cloud-seeding since its infancy."

Although people have been trying to influence the weather since early tribes danced at harvest time, scientific weather modification began after work done in 1946, at the General

ILLUSTRATIONS: K. SIVEYER

Electric Research Laboratory in upstate New York. There, Vincent Schaefer and Irving Langmuir discovered that seeding clouds with carbon dioxide created nuclei around which water could freeze. Bernard Vonnegut (brother of the novelist Kurt) discovered that silver iodide could also be used, a method that has now been adopted across the globe.

Splashing out

Today, countries from Australia to Iran practise some form of cloud-seeding — as do nearly a dozen US states, mainly in the drought-prone western parts of the country. Estimates vary as to how effective the practice is, although it is generally accepted that increases in rainfall of up to 10% can be accomplished in certain types of clouds. California, for instance, claims that its annual spend of around $3 million has generated an additional 370 million to 490 million cubic metres of water a year — a 4% increase over what would happen without cloud-seeding.

Seeding is generally done in one of two ways, depending on the cloud type. In supercooled clouds, which usually reside at high altitudes, water freezes around particles that serve as nuclei. When the ice crystals get too heavy, they fall from the sky and melt as they go, turning into rain or snow. There can be few particles at high altitudes to serve as nuclei, so seeding supercooled clouds — a process called glaciogenic seeding — aims to add more nuclei. Silver iodide is the most widely used glaciogenic chemical, although other materials, such as solid carbon dioxide, can also be used.

Warmer clouds, which are usually at lower altitude, are targeted through hygroscopic seeding. This approach uses compounds such as sodium, lithium and potassium salts. The idea is to generate larger droplets, either by providing larger nuclei to condense around or encouraging small droplets to come together to form larger drops that can fall. The amount of water vapour and size of the water droplets are crucial for the seeding effects.

"Without these cloud-seeding technologies, the Chinese are just shooting in the dark."
— Daniel Rosenfeld

Many other factors affect how well cloud-seeding works, such as when and how to apply the chemicals in question. Another factor is which clouds to target: taking into account their temperature, thickness and convective patterns, and the way that the winds flow into and out of them. The criteria, says Boe, are very stringent, and "the majority of clouds fail".

The best clouds to shoot for, many agree, are orographic clouds, which are produced when air is forced upwards over mountain ranges. Such clouds "are short-lived, relatively shallow and contain much water", says Daniel Rosenfeld, an atmospheric scientist at the Hebrew University of Jerusalem. That's why Weather Modification is targeting orographic clouds in Wyoming for one of the largest US seeding projects, the $9-million, five-year Wyoming Weather Modification Pilot Project.

The Wyoming plan calls for one aircraft and remote-controlled units on the ground, which seed clouds in the winter in an effort to create more snow and build up the snowpack for use in the summer when it melts. The first, ground-based-only seeding took place in the winter of 2006–07; last year, the plane and 25 ground stations were involved. Daniel Breed, a meteorologist at the US National Center for Atmospheric Research in Boulder, Colorado, who is helping evaluate the programme, says "promising results" suggest that seeding materials are getting into the right clouds and showing up as greater numbers of ice nuclei there.

Such work can only help the reputation of weather modification, he says. "There have been a lot of extravagant claims, and it has been a real disservice to science in general, let alone atmospheric science and weather modification."

Each to their own

Most of the work in China focuses on promoting rain and deterring hail, in an effort to reduce damage to agricultural crops, but some of it deals with breaking up fog or diverting lightning[1]. Most of China's 34 districts have their own weather-modification office, and close to two-thirds of its 2,900 counties have their own cloud-seeding stations.

Some Chinese rain-makers have tried to conduct controlled seeding experiments to evaluate how effective the technique is. For example, Shi Li-xin, deputy director of the weather-modification office at the Hebei Meteorological Bureau, and his colleagues have been trying to identify suitable seeding conditions by studying the properties of clouds using a combination of ground and satellite technologies, as well as *in situ* measurements by planes. After assessing

N. H. GUAN/AP

China operates 5,000 rocket launchers as part of its enormous effort to encourage rainfall.

factors such as cloud thickness, the content of supercooled water and the suitable seeding layer of the cloud, the team selected three regions — one operational region of 36,500 square kilometres, and two controls that cover 19,800 and 20,000 square kilometres each. In the early 1990s, the researchers found that rainfall rose by 18% as a result of 21 seeding operations — but the sample was too small for the results to be statistically significant.

In an earlier study, conducted between 1975 and 1986, meteorologists in Fujian province, in southeast China, conducted a randomized seeding experiment with two 14,000-square-kilometre regions. Over the course of 244 experimental days, they found that areas that had been seeded had 20% more rainfall than did those that had been left to their own devices.

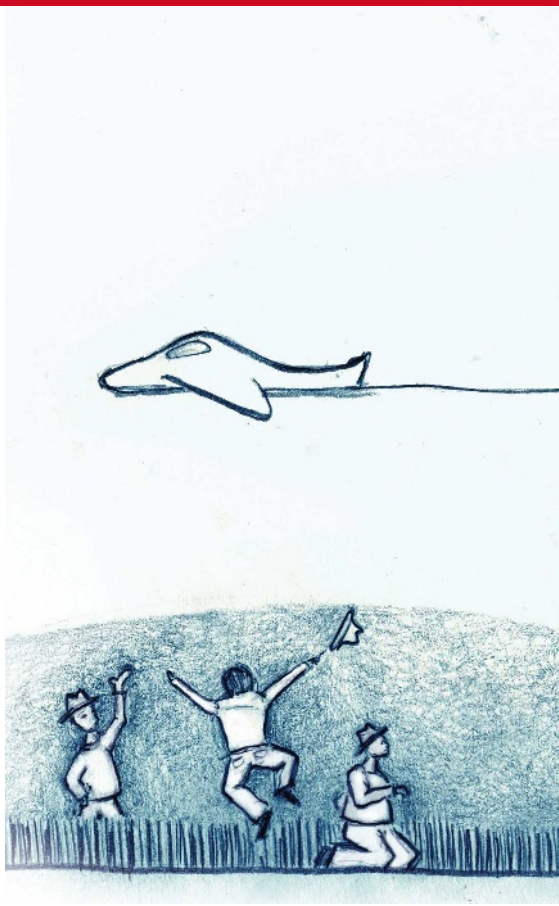
Still, the results of these cloud-seeding experiments have not been published in peer-reviewed journals, and much scepticism lingers. According to Shi, the national statistics on rain creation come from weather-modification offices at all levels, from federal down to provincial and village offices. And as they are tied to performance appraisals and funding for the offices, there is plenty of incentive to exaggerate.

A source close to the China Meteorological Administration, who asked not to be named for fear of political repercussion, says that such scepticism is secretly shared by many Chinese atmospheric scientists and meteorological officials. "You may think that the solution is straightforward: stop applying cloud-seeding technologies until their effectiveness is proved," he says. "But it's not so simple."

"We have been in the dark ages in terms of scientific research in weather modification for going on 25 years."
— William Cotton

"There is a huge demand from the farmers for those technologies," says Lu Da-ren, a researcher at the Beijing-based Institute of Atmospheric Physics, part of the Chinese Academy of Sciences. "So it's not just a scientific issue." Lu admits that the effectiveness of weather modification needs to be better validated, but maintains that this should not exclude a trial-and-error approach at the same time. "Maybe it's not as effective as one thinks," he says. But "from the farmers' point of view, it's better than nothing".

For the Olympics, the Beijing Meteorological Bureau aims to change how and when the rain falls. Starting in 2002, the year after Beijing was chosen to host the games, the bureau has used radar, satellites and weather balloons to scrutinize properties such as structure, temperature and size of droplets of clouds passing over Beijing, as well as over the adjacent Tianjin municipality and Hebei province.



Chinese rain-makers plan to get the clouds to rain out before reaching Beijing, or to prevent small clouds from getting bigger by dissipating them with salt. Failing that, they will over-seed the clouds hoping to reduce the size of each water droplet or ice crystal, which is then

Forget the weather forecast

Lightning

Since Benjamin Franklin's discovery that lightning is a type of naturally occurring electricity, people have tried to find ways to draw it down at specific locations. Those attempts rely on the fact that lightning travels most readily down metals, which have a small electrical resistance, and could help develop and test devices designed to prevent lightning strikes.

Researchers in countries such as China, Brazil and the United States shoot rockets into thunderstorms to divert lightning through a wire or wire–nylon hybrid connected to the ground. Qie Xiu-shu, of the Beijing-based Institute of Atmospheric Physics — known as the 'lightning lady' by the Chinese atmospheric-science community — and her colleagues show that this approach has shed fresh light on the physical processes of lightning[3].

In addition, shining high-power laser light into a storm could eject electrons from molecules in the air. The freed electrons could then act as a conducting wire. There have been some successes in triggering electrical activity in clouds with laser pulses, but the current lasers are still too weak or discontinuous to draw lightning to the ground[4].

Hail

Silver iodide particles sprayed into clouds are supposed to not only promote rain, but also reduce the size of damaging hailstones. Theory holds that creating more droplets spreads the available water more thinly between them, thus reducing the number that grow to hail size.

Farmers' insurance claims for crop damage have dropped dramatically in some places where weather-modification experiments have been done. One 1997 study in North Dakota showed a significant reduction in payments to farmers for hail damage in years when hail-suppression technologies were used[5].

Hurricanes

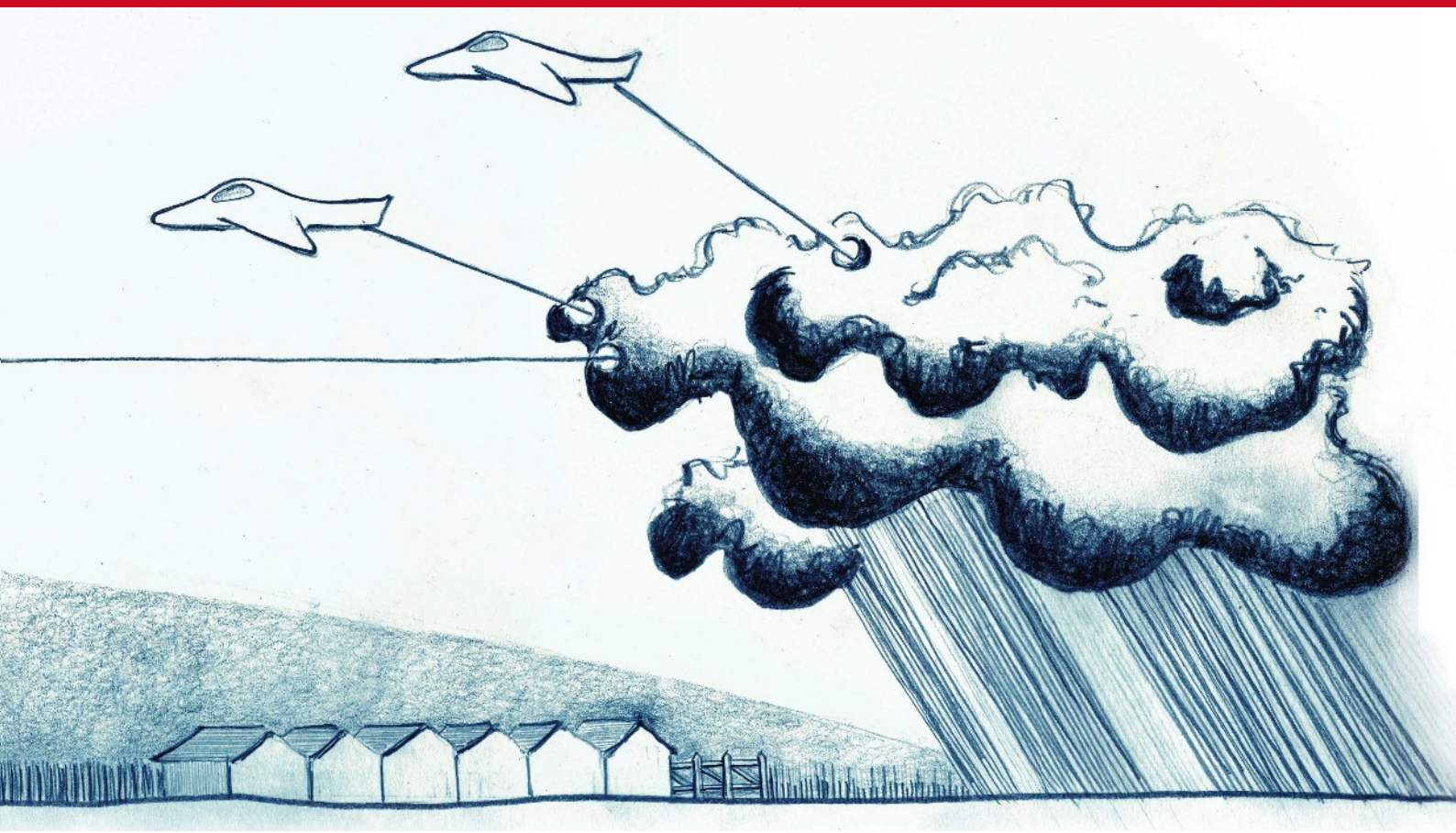
America's Project Stormfury, run between 1962 and 1983, never succeeded in its goal of weakening hurricanes as they approached land. It used silver iodide to seed clouds outside the hurricane, in the hope that this would lead to a larger eyewall and therefore lower overall wind speeds.

More recently, Daniel Rosenfeld, an atmospheric scientist at the Hebrew University of Jerusalem, has suggested that hurricanes' fury might be mitigated by damping down the production of warm rain at their edges[6]. Similarly, simulations run by William Cotton at Colorado State University in Fort Collins and his colleagues seem to suggest that dumping dust or other particles into a hurricane might, in some cases, reduce the storm's force[7].

Fog

Officials at several airports worldwide use ice-seeding agents to disperse fog and improve visibility on runways. The well-established technique seems to work only with supercooled fog, which is close to freezing.

Unlike other ways to modify the weather, the effects of fog dispersal are easy to see: "You can see it clearly because where you see it, you create holes in the fog," says Rosenfeld. D.C. & J.Q.



more likely to be dissipated before reaching the ground. The meteorological bureaus have amassed cannons, rocket launchers and planes in more than 100 locations in Beijing, Tianjin and Hebei.

In recent months, the bureaus have intensified their field testing. The Beijing Meteorological Bureau declined to provide details of the results, but maintains that keeping the opening ceremonies dry will depend largely on the cloud properties and the accuracy of the weather forecast. "We can achieve reasonably good results with local, weak weather patterns, but are unable to bring about complete elimination of rain in face of large, thick clouds covering a large region," the bureau said in a statement.

Sketchy evidence

Part of the problem is that so little is known about the mechanisms through which cloud-seeding might work. An influential 2003 report[2] from the US National Research Council (NRC) declared that "there still is no convincing scientific proof of the efficacy of intentional weather-modification efforts". A riposte from the Weather Modification Association accused the members of the panel of lacking experience or knowledge of weather modification — and said they had held cloud seeding to a standard of scientific proof "that few atmospheric problems could satisfy".

For one thing, clouds can be very different at distinct locations; they also vary over time at a particular place. And lack of funding in some countries hasn't helped; in the United States, for example, federal funding for weather modification peaked at around $20 million a year in the late 1970s, but is now negligible. States have been left on their own to fund individual projects. "We have been in the dark ages in terms of scientific research in weather modification for going on 25 years," says William Cotton, an atmospheric physicist at Colorado State University in Fort Collins. "And so when somebody wants to have something really quantitative, we can't deliver."

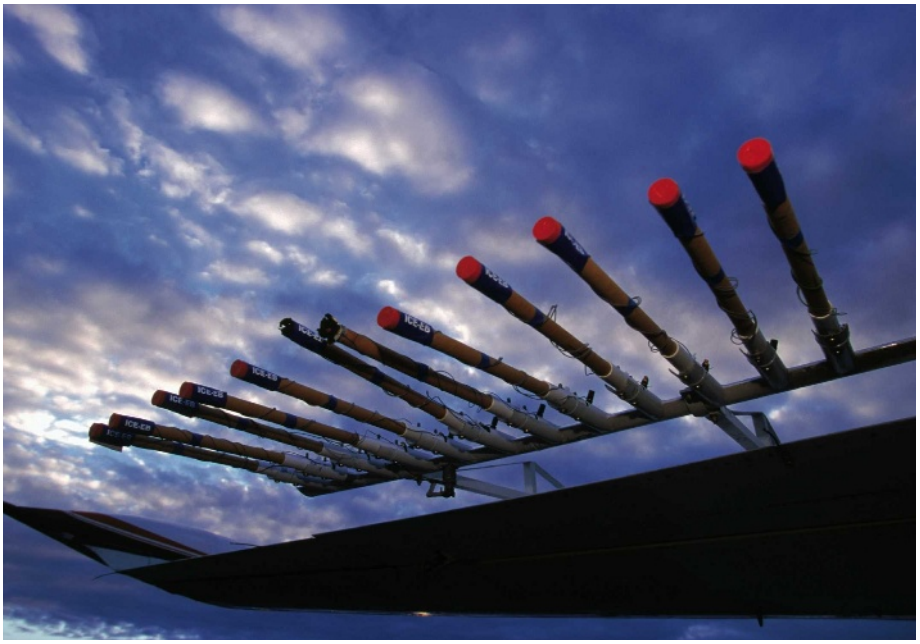
"It is the same problem as weather forecasting," adds Deon Terblanche, divisional manager of research for the South African Weather Service in Pretoria and chairman of the World Meteorological Organization's Expert Team on Weather Modification. "It's very difficult to forecast even 30 minutes ahead what a specific cloud will do. If you do something to that cloud and you can't even forecast what it will do exactly in nature, it becomes difficult to prove what your effect was." He does, however, think that some of the weather-modification technologies currently in use show a big enough effect to be statistically significant.

Choosing which cloud to seed, and which approach to use, can greatly affect the outcome.

WEATHER MODIFICATION CENTRE, CHINA METEOROLOGICAL ADMINISTRATION

Michael Manton, of Monash University in Victoria, Australia, cites three reasons for the difficulties in proving the impact of cloud seeding: a mismatch between the scale of the impact and the scale at which seeding acts; the natural variability of rainfall versus the incremental impact of seeding; and the expense of evaluation. "I believe we have not moved significantly forward since [the NRC report in] 2003," he says. "In many parts of the world there has been an implicit or explicit faith in cloud seeding, so seeding has not been carried out under carefully controlled conditions."

Moreover, there is growing realization of the importance of air-pollution aerosols on cloud formation and precipitation. Some researchers think that aerosols can reduce rainfall by decreasing the size of water droplets in warm clouds; there are also indications that aerosols could affect precipitation in cold clouds and the dynamics between the two cloud systems, says Zhanqing Li, of the University of Maryland in College Park. "This may be a missing piece of the jigsaw in the puzzle of cloud-seeding


P. MCBRIDE/AURORA/GETTY IMAGES

Light aircraft are used to seed clouds with compounds such as silver iodide.

research. Under inappropriate seeding conditions, we may get the opposite effect to what is intended."

The problem with aerosols underlies a new debate on whether Israel's randomized seeding experiments — once seen as some of the most convincing evidence for the effects of cloud-seeding — were as effective as once thought. Two long-running experiments, each of which was carried out over 6 years in the 1960s and 1970s, suggested that seeding increased rainfall by 12–15%. But some have questioned those conclusions. Most recently, an ongoing study led by Zev Levin of Tel Aviv University took aerosols into consideration and showed that seeding might have zero or even negative effects in regions with high levels of pollution. "This is just the beginning of the debate, but strikes a note of caution on large-scale cloud-seeding operations in the presence of heavy pollution," says Li. This is, of course, a particular concern for China, with its heavy aerosol burden.

Israel's water commission has just launched another major cloud-seeding experiment, with a budget of US$1 million a year for several years. Researchers from the Hebrew University of Jerusalem and Tel Aviv University will build on what has been learned from previous experiments and incorporate cutting-edge technologies to measure cloud properties and to trace seeding materials and assess their role in rain formation. The ultimate goal is to see whether technologies can replenish water reserves in Lake Tiberius (also known as the Sea of Galilee) in northern Israel.

A report, commissioned by the World Meteorological Organization and the International Union of Geodesy and Geophysics, calls for a better understanding of the role of aerosols in precipitation and climate systems. "Aerosols are involved in a long chain of reactions and complicated feedback mechanisms leading to precipitation," says Levin, chief editor of the report, which will be released within the next month. "Cloud seeding adds to that complexity we know so little of." In addition, the report stresses the serious limitation of using statistical tools in cloud seeding without a proper understanding of the underlying physical processes.

Rosenfeld maintains that the Chinese will be able to make genuine progress only by combining approaches such as measurements of cloud properties, numerical modelling, as well as randomized and targeted seeding experiments. "Without these cloud-seeding technologies, the Chinese are just shooting in the dark," he says.

But long-term, randomized or targeted seeding experiments over a large region would mean forgoing opportunities to seed suitable clouds, and to some Chinese meteorologists, this is unthinkable given the country's water crisis. "If suitable clouds are there, we have to conduct cloud-seeding operations," says Shi. "The farmers would be furious if we didn't." The stakes are so high that in some instances, farmers have accused those in neighbouring villages of "stealing" their rain by seeding passing clouds.

Many researchers are concerned that China's weather-modification scheme has been heavily tilted towards operations, and that researchers and operations managers need to collaborate much more than they have until now. The newly established Centre for Weather Modification at the Chinese Academy of Meteorological Sciences may help start to address some of these issues, says its director, Guo Xueliang. The centre, set up last December, will eventually house some 60 researchers. They will use numerical models, laboratory simulation and field experiments to analyse cloud properties and precipitation principles. "We recognize the importance of basic research in guiding cloud-seeding operations, and will apply scientific rigour in testing their effectiveness in controlled, cross-province field experiments," Guo says.

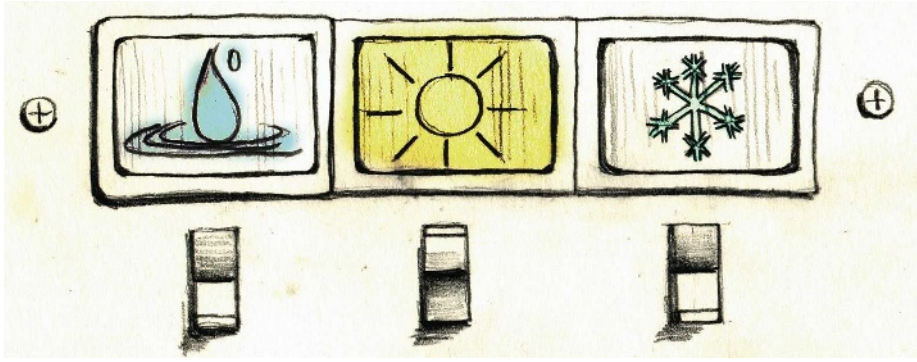
And China's lavish programme is likely to receive even more money; the country has listed weather modification as one of the key projects in its eleventh five-year plan, and says it will introduce the latest equipment for both research and operations. Several large-scale national projects are also in place to aid collaboration between academic institutes and meteorological offices, such as a five-year project funded by the science ministry.

China has found plenty of reasons to move forwards with its work. As the momentum of the Olympics gathers pace, the Beijing Meteorological Bureau is under increasing pressure to give its best performance yet. The promised extravaganza of the opening ceremony will not be the only focal point as the opening ceremonies begin; meteorologists around the world are also eager to hear what Beijing will say about its rain-suppression operations — especially if the weather turns out to be dry. ■

Jane Qiu writes for *Nature* from Beijing, and Daniel Cressey from London.

1. Guo, X. & Zheng, G. *Adv. Atmos. Sci.* (in the press).
2. National Research Council. *Critical Issues in Weather Modification Research* (National Academies, 2003).
3. Yang, J. *et al. Atmos. Res.* (in the press).
4. Kasparian, J. *et al. Opt. Express* **16,** 5757–5763 (2008).
5. Smith, P. L., Johnson, L. R. & Priegnitz, D. L., Boe, B. L. & Mielke, P. W. *J. Appl. Meteorol.* **36,** 463–473 (1997).
6. Rosenfeld, D. *et al. Atmos. Chem. Phys. Discuss.* **7,** 5647–5674 (2007).
7. Cotton, W. R., Zhang, H., McFarquhar, G. M. & Saleeby, S. M. *J. Weather Modification* **39,** 70–73 (2007).

See Editorial, page 957.



The research revolution

As the first grants from the European Research Council begin to come through, **Geoff Brumfiel** investigates whether the new system is meeting its goals.

Markus Reichstein is obsessed with dirt. If he could just do a better job of simulating the stuff, says the 35-year-old climate modeller, he could minimize a major source of uncertainty in climate predictions. His fellow climate modellers "don't like to dig into the soil", Reichstein explains from his office at the Max Planck Institute for Biogeochemistry in Jena, Germany, located in the hills above that city's medieval streets and Communist-era smokestacks. Instead, they represent Earth's incredibly complex and dynamic top layer — which is a huge carbon reservoir — with ridiculously simple approximations.

Until recently, however, Reichstein's obsession with improving this situation was frustrated by a shortage of money. Research funding in Europe varies considerably from country to country, but scientists are typically funded through their home institutions. And the level of funding rises with seniority — which is why senior researchers receive the bulk of the funding and run the majority of groups. Reichstein wasn't old enough to get all the money he needed from the Max Planck Institute. Nor was he in any position to go after one of the big research grants given out by the European Union (EU). Those awards are targeted at multinational collaborations, with a stated goal of strengthening ties between member nations. Besides, says Reichstein, who was as frustrated by the system as most other young scientists in Europe, the resulting collaborations "went too much in the direction of applied science". There was no obvious place for him to marshal the resources needed to tackle his big questions about dirt.

That is why Reichstein paid close attention last year when the European Commission launched the European Research Council (ERC): a semi-autonomous agency that would award its grants in a decidedly different way. Instead of focusing on political goals, its only criterion would be the quality of scientific proposals as judged by an international group of peer-reviewers. By European standards "we are absolute radicals", says Fotis Kafatos, the ERC's president and an immunogeneticist at Imperial College in London. "There is no consideration of nationality in the evaluation — absolutely none."

G. BRUMFIEL

Markus Reichstein's grant will enable him to improve the simple approximations of soil used in climate models.

"We are absolute radicals."
— Fotis Kafatos

The impetus for this radical idea came from the scientific community itself, starting around the turn of the millennium. Researchers were increasingly concerned that the politically oriented selection process of the EU was overlooking some of the best science. Their model was the US National Science Foundation (NSF), a US$6-billion agency that has been making peer-reviewed grants to individual researchers (as well as to its larger centres) for nearly 60 years, and that is regularly praised for its ability to fund the best research in a broad range of fields. "The dream of everyone is the National Science Foundation," says Ernst-Ludwig Winnacker, the ERC's secretary general, based in Brussels, Belgium.

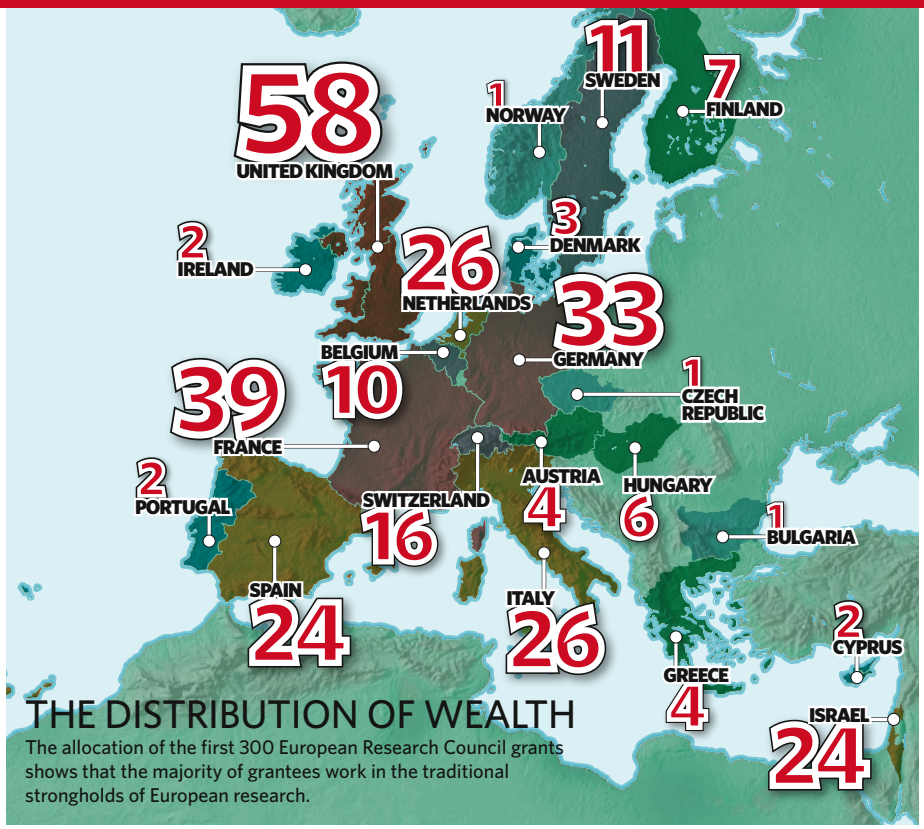
The ERC was formed in February 2007 as part of the EU's Seventh Framework Programme, which sets the research funding trajectory from 2007 through 2013. The ERC's total budget for that period is €7.5 billion ($11.6 billion), or about €1 billion a year. And it mirrors the NSF in several ways: its operations are overseen in part by a scientific council independent of the EU itself; it is divided into directorates that cover everything from the social sciences to physics; and most importantly, it makes its grant selections using a continent-wide network of peer reviewers.

Reichstein was particularly taken with the new council's emphasis on fundamental research throughout Europe. National funding bodies can be parochial in their choice of which projects to back, he says. The ERC brings a broader perspective. Better still, the ERC's first round of granting would target researchers like him: early-career scientists who had completed their PhDs between two and nine years ago, and were in the process of establishing an independent group.

Flood of applications

Reichstein saw the ERC as a perfect opportunity to dig into dirt, and lost no time in applying. He was hardly alone: the council received almost 9,200 applications for just a few hundred grants, according to Helga Nowotny, a social scientist from ETH Zurich (the Swiss Federal Institute of Technology) and member of the ERC board. "Some panels were just overwhelmed by sheer numbers," she says. To deal with the huge volume of applicants, extra evaluators had to be brought in, and a prescreening process was set up.

The process was further strained by the bureaucratic rules of the European Commission, according to Robert May, a zoologist at the University of Oxford, UK, and, until he stepped down on 31 May, a member of the ERC



SOURCE: ERC

THE DISTRIBUTION OF WEALTH

The allocation of the first 300 European Research Council grants shows that the majority of grantees work in the traditional strongholds of European research.

scientific council. Peer-reviewers were required to fill out lengthy conflict-of-interest forms and board members were saddled with travel regulations. Although no fault of the ERC directly, he says, the granting system "is hedged about with multiple bits of paper that makes everything extravagantly cumbersome".

The selection process

Still, the grant-allocation process moved forward. The pre-screening reduced the pool of applicants to about 550 investigators, who were then brought to Brussels and interviewed. Reichstein was one of them. After a five-minute presentation about his project, reviewers grilled him with questions that were "critical but very constructive", he says. "They wanted to find out if I was really dedicated."

The ERC finally announced its first round of grants in December 2007, relatively on schedule. Reichstein heard that he'd been selected while he was attending the American Geophysical Union's annual meeting in San Francisco. He instantly brought a round of beers for his colleagues. "It was surreal at that point," he says.

Over the next five years he will receive €1 million to integrate soil data from European observation stations into global climate models. The ambitious project will seek to characterize microbes in the soil, understand carbon transport through its layers, and ultimately develop computer code that can replace current 'black box' models of dirt. "The final goal is to move towards a more realistic description of the soil," he says. "And I think we are at the point where we can."

Other first-round winners are similarly enthused about the opportunities the grants provide. "In Italy it's very difficult for young researchers," says Livia Conti, a physicist at the National Institute for Nuclear Physics in Padova. Conti will use the money to look at how temperature fluctuations can contribute to noise in gravitational-wave detectors. She says that the funds will allow her to hire a theorist.

"This is way bigger money than anything you could get in Sweden," agrees Johan Elf, a molecular biologist at Uppsala University. Elf is studying single-molecule dynamics inside cells, and the money will go towards buying specialized equipment and hiring more staff. In September, Elf returned to Sweden after a two-year postdoc at Harvard University. "The ERC money is a great motivation for staying in Europe," he says.

> "The grant allows me to attack a big question. That would not have been possible before."
> — Markus Reichstein

Of course, the old bureaucratic barriers haven't vanished overnight. Red tape is still holding up funding for some, although Reichstein's contracts have finally been signed. The general feeling is that the process was successful, says Nowotny: "Overall it went surprisingly well."

Members of the scientific council are especially relieved that there has been so little political resistance to the ERC, even though the vast majority of winners work in the traditional strongholds of European research (see map). The United Kingdom, France and Germany together are home to nearly half of the selected proposals. Newcomers to the EU such as the Czech Republic and Hungary fare much worse, hosting only a handful of winners between them. Italian scientists had the lowest success rate — submitting more than 1,500 applications but winning just 26 grants.

Nonetheless "neither the [European] Commission nor the Parliament has interfered with the process," says Winnacker. "I actually expected more problems from the politicians," adds Michał Kleiber, a computer scientist at the Polish Academy of Sciences in Warsaw and a member of the ERC science council. "To my pleasant disappointment, there wasn't much heard from Poland." Kafatos thinks that countries on the losing end of the first round didn't object because they saw the ERC's process as fundamentally transparent and fair. "They are disappointed," he says. "But they also realize that they have to do some homework."

With the first round of grants now complete, the ERC is looking to the future. Later this year, they will award a separate round of 300 'advanced grants' for senior researchers. Meanwhile, they are in the process of nearly doubling their staff of 110 and moving towards becoming a European Commission 'executive agency', which should allow them to issue grants more quickly with less paperwork.

The council is also working to reduce the number of substandard applications by requiring applicants to demonstrate a track record and supply a five-page technical summary of their work. Additionally, the eligibility period for young investigators will be narrowed to just three to eight years after their PhD.

Even then, not all the best applications will receive funding. This year, for example, some 130 applicants passed the agency's threshold, but did not receive money. Kafatos would like national governments to help: "We would like to see the national system use the results in ways that might be helpful to them," he says. And that is beginning to happen: France, Italy, Spain and Switzerland have begun national initiatives to fund young investigator grantees that the ERC ranked highly but was not able to fund.

Overall, the ERC is off to a running start, says May. "The critical hurdles have been cleared," he says. Kafatos agrees: "It's not everyday that you get more than 9,000 applications for a first call. It has been an amazing experience." Indeed, says Kafatos, the Council has seen a more manageable volume of applicants for the grants going to senior researchers, thanks in part to refined requirements.

Back in his office in Jena, Reichstein is gearing up for his grant, which will begin in September. He says he will use the money to fund two PhD students, a postdoc and an assistant to begin working on the improved soil model. "It really allows me to attack a big question," he says. "That would not have been possible before." ■

Geoff Brumfiel is a senior reporter based in London.

See Editorial, page 958.

CORRESPONDENCE

Fewer academics could be the answer to insufficient grants

SIR — The rejection of high-quality grant proposals is a problem endemic to universities throughout the world. I suggest that it arises from separating the employment of academics from the central bodies who provide grant funding.

Consider the country of Euphoria. It has just four universities, each of which employs ten academics of comparable quality, and one national funding body. Each academic submits two grants per year, only one of which is rated fundable. So each academic is awarded one grant per year. Into this happy state enters the ambitious new president of the Euphoric University of Fulchester. He makes his mark by doubling the number of academics in his institution. Now Euphoria has 100 grants submitted per year, but still only 40 grants available. Fulchester will get 16 of these and the other universities will now get only eight each.

This is a rational action by the new president, as the rewards from obtaining six additional grants are so great that it is worth hiring the 10 new staff. Euphoria loses overall, however, because its taxpayers and students are now paying to employ 10 extra staff, with the same amount of research being done. The other four universities also lose, as they are now receiving two fewer grants. The incentive will be for them to act in a similar way until Euphoria stabilizes, with many more excellent grants being submitted than can be funded.

This situation naturally arises in an environment in which employing academic staff is separated from obtaining research funding. We would be better off having fewer academics and using the savings to fund more grants, because then more research could be done for the same national expenditure. Such action has to be taken by governments, as universities currently have the freedom to over-staff and are rewarded for doing so under the present system.

Andrew Doig Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

Working together to put molecules on the map

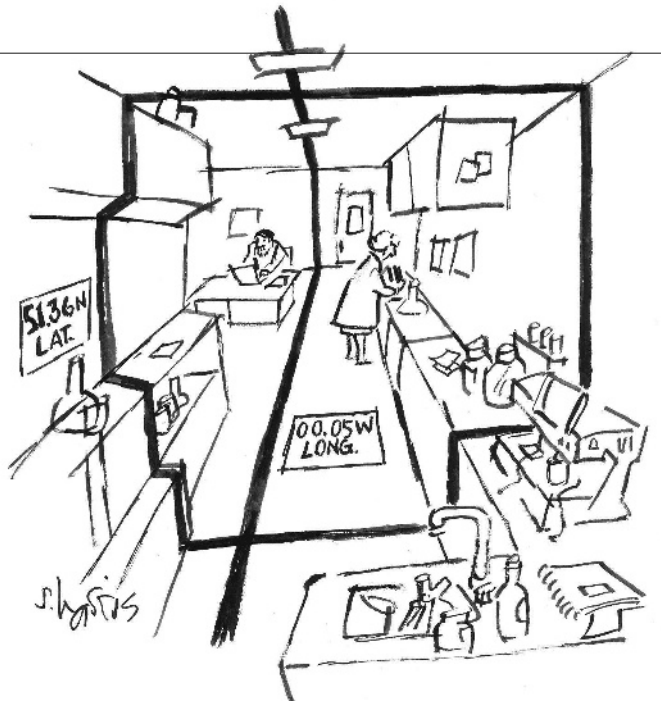
SIR — We applaud the call in your Editorial 'A place for everything' (*Nature* **453,** 2; 2008) for researchers to record the latitude and longitude of their data, in order to place all biological samples in proper spatial (and temporal) context. We agree that this minimum information guideline should apply to all biological samples taken from the natural environment, and note the pressing need for relevant molecular data to be tagged with geographical location.

The International Nucleotide Sequence Database Collaboration — comprising the DNA Data Bank of Japan, the European Molecular Biology Laboratory and GenBank — already offers the option of recording latitude and longitude coordinates. This qualifier, among others, was requested by the Consortium for the Barcode of Life to provide the geographical origin of molecules it uses to identify organisms. The "minimum information about a genome sequence" guideline published by the Genomic Standards Consortium (*Nature Biotech.* **26,** 541–547; 2008) calls for this critical field to be mandatory for all genome and metagenome submissions, along with altitude or depth and time of sampling.

Other molecules that are equally critical to tag with this information are the vast number of other marker genes, especially 16S and 18S ribosomal RNA sequences, that are being generated globally from a diverse range of habitats. This registration becomes all the more relevant as ultra-high-throughput sequencing of these molecules continues to be more widely applied. Core to these efforts are projects such as the Environment Ontology and Gazetteer initiatives, which describe environments and place names, respectively. Combined, these resources will support the consistent annotation and retrieval of environmental information associated with an organism or biological sample.

These projects all highlight the growing importance of community-driven initiatives in developing improved standards for reporting experimental data. We look forward to the day when it will be commonplace to view collections of molecules 'on the map', so to speak, such that questions relating to their global and local abundances, distributions, environments and functions can be properly addressed. Getting to this point will require: increased awareness; higher expectations for the quality and quantity of descriptive data recorded; improved standards, ontologies and databases; proof of the value of downstream analyses; and widespread practical changes, such as use of hand-held devices for recording real-time contextual information (and, in the future, for generating data) in the field.

Dawn Field NERC Centre for Ecology and Hydrology, Oxford OX1 3SR, UK
This letter was also signed by the following, whose addresses can be found at http://gensc.org:
Norman Morrison, Frank Oliver Glöckner, Renzo Kottmann, Guy Cochrane, Robert Vaughan, George Garrity, Jim Cole, Lynette Hirschman, Lynn Schriml, Ilene Mizrachi, Scott Federhen, David Schindel, Scott Miller, Paul Hebert, Sujeevan Ratnasingham, Robert Hanner, Linda Amaral-Zettler, Mitchell Sogin, Michael Ashburner, Suzanna Lewis, Barry Smith, Genomic Standards Consortium (GSC), International Nucleotide Sequence Database Collaboration (INSDC), Consortium for the Barcode of Life (CBOL), International Census of Marine Microbes (ICoMM), Environment Ontology Consortium (EnvO)



Decoherence does not get rid of the quantum paradox

SIR — In his Essay 'Lifting the fog from the north' (*Nature* **453,** 39; 2008), Maximilian Schlosshauer describes how the process of

"The undeciphered Phaistos Disc is perhaps the most infamous of ancient inscriptions." Andrew Robinson, page 990

decoherence can explain the famous double-slit experiment. An electron interacting with innumerable quanta in the photographic plate (and its environment) becomes entangled with all of them — and the resulting collective wavefunction is so narrow that it appears particle-like.

But the question remains as to why the wavefunction narrows in precisely the location where it does, or — as Schlosshauer puts it — "Why is a single spot here and not there?"

The author's somewhat 'foggy' answer is suggestive of a version of Everett's 'many worlds' idea (see *Nature* **448,** 15–17; 2007), in which all possible branches of the wavefunction continue to exist autonomously. But this interpretation merely shifts the question to "Why do I find myself experiencing the branch/world with the spot here and not the branch/world with the spot there?"

We still have no answer and, if there is one, decoherence is at best only part of it (S. L. Adler *Stud. Hist. Philos. Sci.* **34,** 135–142; 2003). As Joos and Zeh remarked on decoherence as a source of spatial localization: "Of course no unitary treatment of the time dependence can explain why only one of these dynamically independent components is experienced." (E. Joos and H. D. Zeh *Zeitschrift Phys. B* **59,** 223–243; 1985).

We are still left with a dichotomy: on the one hand, infinitely many continuously distributed potentialities, and on the other, one narrow, irreversibly realized actuality. Contrary to Schlosshauer's conclusions, complementary (mutually incompatible) descriptions are necessary to describe the landscape we are currently experiencing, even as the fog is lifting.

Nikolaus von Stillfried
Department of Environmental Health Science, University of Freiburg, Breisacherstrasse 115B, 79106 Freiburg, Germany

Ventures should not overstate their aims just to secure funding

SIR — Revolutions are often conceived with the best intentions, but so easily claim more than is plausible and more than can ever be delivered. We fear that the "revolution in climate prediction" called for by the World Modelling Summit for Climate Prediction and reported in your journal ('They say they want a revolution' *Nature* **453,** 268–269; 2008) will fall foul of the same hubris.

Any venture bidding for investment that exceeds a billion dollars needs to have well-grounded justifications. Advancing our basic understanding of how the climate system works through enhanced representation of that system in next-generation climate models — ("pure intellectual excitement") — may indeed be such justification. But claiming that this will allow scientists to "provide answers to key questions ... such as future food supply" and guide decisions the world will be making to cope with climate change displays a misunderstanding of the nature of adaptation and its contingency on our imagining of future social change.

The reason that the UK summit at Reading University over-claimed the benefits of climate prediction for adaptation in its pitch for a billion dollars of new science investment is revealed by the summit's chair, Jagadish Shukla, in his warning: "If we just ask for enhanced understanding, then we have very little chance of getting the necessary funding".

Effective and robust adaptation strategies are not significantly limited by the absence of accurate and precise regional climate predictions. They are limited more by a multitude of technological, institutional, cultural, economic and psychological factors that lie beyond the reach of climate models — and always will. The epistemological limits to predicting future climates with accuracy and precision must not be used as a reason to limit adaptation to climate change. Bring on the revolution if you will, but don't mistake it for Utopia.

Mike Hulme, Suraje Dessai Tyndall Centre, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Digital identifiers work for articles, so why not for authors?

SIR — Several Correspondences, including 'Give south Indian authors their true names' (*Nature* **452,** 530; 2008) and 'Name variations can hit citation rankings' (*Nature* **453,** 450; 2008), have illustrated difficulties in identifying authors and their papers, citations and *h*-index.

In an academic world in which decisions on promotion and funding often depend on the applicant's scientific impact, an incorrect publication or citation record in an online database can be very inconvenient. Scopus and Thomson's Web of Science, which make available abstract and citation databases, acknowledge the issue and have come up with solutions: the Author Identifier and ResearcherID, respectively.

These systems assign an identifying code to each author. Unfortunately, a single author can have more than one Author Identifier in Scopus (I am cryptically known as 7006716603 and 16551750300). And as only invited researchers can register for a number, ResearcherID is not yet used as a unique author key in the Web of Science — making it difficult to differentiate me from a highly cited ecologist from the Netherlands, despite the 'Distinct Author Sets' feature.

If it is possible to have DOIs for objects (or, so they say, enough IPv6 addresses for every molecule on Earth), why is it so difficult to implement DAIs for authors?

Raf Aerts Division Forest, Nature and Landscape, Katholieke Universiteit Leuven, Celestijnenlaan 200E-2411, 3001 Leuven, Belgium

Europe needs to protect its transgenic crop research

SIR — On 5 June 2008, our authorized, small-scale field trial of transgenic potato plants for nematode control was destroyed by people seeking to coerce government and society. It was one of only two trials authorized in the United Kingdom this year.

Our concern is that Directive 2001/18/EC, the European Union (EU) legislation that governs such trials, is confused. Although it recognizes the need for field releases at the research stage (clause 23), it does not distinguish between these and development-stage trials in its risk assessments. It has also set the legal precedent of providing precise locations of trial sites to vandals.

We have no evidence that the 400 transgenic plants we released posed any environmental concern, particularly when considered in the context of the annual UK potato crop of 8,000 million plants and their naturally hazardous glycoalkaloid content.

If EU governments cannot protect the trials they authorize, they should establish secure, vandal-proof national testing centres.

Unfortunately, a failure to distinguish a research trial from product-development trials seems to have blinded activists to the published, broader aims of our work. We develop controls for nematodes on subsistence crops in Africa and Asia, where both farmers and governments recognize the need for new technologies.

What is the distinction between burning university books 75 years ago and now destroying university research intended for publication in scientific journals? European governments must ensure that science in our universities can progress without coercion.

Howard J. Atkinson, Peter E. Urwin Centre for Plant Sciences, University of Leeds, Leeds LS2 9JT, UK

COMMENTARY

Repairing research integrity

A survey suggests that many research misconduct incidents in the United States go unreported to the Office of Research Integrity. **Sandra L. Titus**, **James A. Wells** and **Lawrence J. Rhoades** say it's time to change that.

Misconduct jeopardizes the good name of any institution. Inevitably, the way in which research misconduct is policed and corrected reflects the integrity of the whole enterprise of science. The US National Academy of Sciences has asserted that scientists share an 'obligation to act' when suspected research misconduct is observed[1]. But it has been unclear how well scientists are meeting that obligation. In the United States, the Office of Research Integrity (ORI) evaluates all the investigation records submitted by institutions and plays an oversight role in determining whether there has been misconduct at institutions that receive support from the Department of Health and Human Services (DHHS). The reported number of investigations submitted to ORI has remained low: on average 24 institutional investigation reports per year[2].

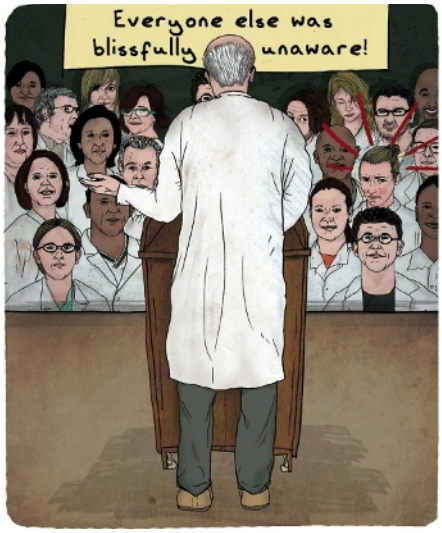
ORI focuses resources, not only on evaluating institutional reports of research misconduct but also on preventing misconduct and promoting research integrity through deterrence and education. To evaluate these initiatives, we investigated whether the low number of misconduct cases reported to ORI is an accurate reflection of misconduct incidence, or the tip of a much larger iceberg. The latter seems to be the case.

The 2,212 researchers we surveyed observed 201 instances of likely misconduct over a three-year period. That's 3 incidents per 100 researchers per year. A conservative extrapolation from our findings to all DHHS-funded researchers predicts that more than 2,300 observations of potential misconduct are made every year. Not all are being reported to universities and few of these are being reported to the ORI.

No regulatory office can hope to catch all research misconduct and we think that the primary deterrent must be at the institutional level. Institutions must establish the culture that promotes safeguards for whistleblowers and establishes zero tolerance both for those who commit misconduct and for those who turn a blind eye to it.

Defining misconduct

A first step in developing that culture is taking stock of misconduct's frequency. Several investigators have addressed research misconduct incidence with limited results because of methodological problems, such as applying inconsistent definitions of misconduct or not accounting for duplicate reports of the same incident[3–5]. So, we used the US federal definition of research misconduct[6] — fabrication, falsification or plagiarism in proposing, performing or reviewing research, or in reporting research results — and verified whether reports accurately fitted that definition. The possibility of duplicate reports was virtually eliminated by selecting only one National Institutes of Health (NIH)-funded researcher in a given department to respond. We asked about events only from the past three academic years to avoid inclusion of distant events and to have a consistent time parameter. We used frequent and varied reminders to secure a high response rate to the survey. Previous research has treated survey reports of misconduct as if the observer could make the determination that they had observed misconduct. Instead, we consider the observations to be 'possible research misconduct' and not all such observations will result in a finding of misconduct. In all we asked 4,298 scientists holding NIH extramural research funds at 605 institutions to respond to the survey so that our findings would be representative of a broad spectrum of research fields as well as varied sizes of institutions.



ILLUSTRATIONS BY J. TAYLOR

"Institutions must establish safeguards for whistleblowers."

What scientists saw

In 2006, we asked participants to indicate the number of times they had observed suspected research misconduct in their own department in the past three academic years (2002–05). 2,212 scientists provided complete responses to questions concerning research misconduct (51% response rate). Of these, 192 scientists (8.7%) indicated that they had observed or had direct evidence of researchers in their own department committing one or more incidents of suspected research misconduct over the past three academic years. The 192 scientists described a total of 265 incidents.

Scientists were asked to indicate how they became aware of the possible misconduct and were told to report observations and not hearsay (see table, page 982). Suspected misconduct was observed at all scientific ranks including postdocs, students, and tenured faculty members. The following are examples of how scientists described such incidents. We used these descriptions to validate whether the observation met the federal definition of research misconduct.

"A post doc changed the numbers in assays in order to 'improve' the data."

"A colleague duplicated results between three different papers but differently labelled data in each paper."

"A co-investigator on a large, interdisciplinary grant application reported that a postdoctoral fellow in his laboratory falsified data submitted as preliminary data in the grant. As principal investigator of the grant, I submitted

supplementary data to correct the application."

"A colleague used Photoshop to eliminate background bands on a western blot to make the data look more specific than they were."

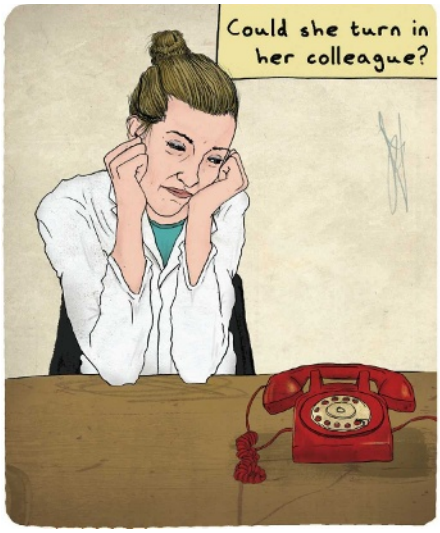
Two people independently coded and evaluated the 265 descriptions to determine whether each met the federal definition of research misconduct. In all, 64 reports (24% of the total) did not meet the threshold of the federal definition — which left 201 observations of potential misconduct made by 164 scientists (7.4%). These 201 misconduct observations included fabrication or falsification (60%) and plagiarism only (36%).

According to our respondents, 58% of the observed incidents had been reported to officials at their institutions. In 24% of incidents it was the survey respondent who reported it and in 33% of the incidents it was someone other than the respondent. Responses indicated that 37% of incidents were not reported by anyone and for 5% of the cases respondents did not know.

Study limitations

Several limitations may have affected the study results. As the sample only includes one observer per department, the number of suspected research-misconduct incidents found in this study is likely to be a very conservative estimate. Because the sample only represented scientists holding research awards given to established researchers, we lack the views of postdoctoral fellows, graduate students, clinical-trial coordinators and lab technicians who might report a different quantity and type of suspected research misconduct. The study is also probably more representative of the biomedical, behavioural and life sciences than it is of the physical and social sciences, reflecting the mission of the NIH.

Although the scientists we sampled were receiving research support from the NIH, we know nothing about the funding of those they suspected to be committing misconduct. This means that the findings do not exclusively apply to NIH investigators. And because of the possibility of human error from respondents, our method of measurement may have failed to elicit all instances of suspected research misconduct or may have included erroneous instances. Some observations, for example, may have occurred outside the time period specified because of 'telescoping' — including salient events that occurred before the period of interest. Still, the questionnaire was careful to specify the period of interest as the past three academic years.



Extrapolating the survey results — even conservatively — projects an alarming picture of under-reporting. NIH extramural research grants in 2007 supported an estimated 155,000 people, which includes principal investigators and other research personnel[7]. In our survey, 201 cases were observed over three years by 2,212 respondents, essentially 3 cases per 100 people per year. Most conservatively, we assumed that non-responders (roughly half of our sample) did not witness any misconduct. Thus, applying 1.5 cases in 100 scientists to 155,000 researchers suggests that there could be, minimally, 2,325 possible research misconduct observations in a year. If 58% of these cases were reported to institutional officials as in our survey, approximately 1,350 would have been reported whereas almost 1,000 could be assumed to go unreported to any official.

"Extrapolating the survey results projects an alarming picture of under-reporting."

These numbers indicate a sizeable disconnect between what universities are seeing and the 24 investigations evaluated by the ORI annually. Could all the predicted cases be found to lack evidence? Could all the cases be concluded at the inquiry stage? Could the cases be primarily occurring in research that is not funded by the Public Health Service and hence not reportable to the ORI? Can duplicate observations of misconduct account for this disparity?

We doubt that affirmative answers to these questions could sufficiently explain the discrepancy. We recognize that this estimate is not perfect. First we are applying our findings from a defined context to a much larger context and one that also includes the staff of the investigator. Another weakness of the prediction is the fact that scientists in our study would have been narrowly reporting observations restricted to their own experience. A single observer in a department cannot be expected to have been exposed to all instances of misconduct. Thus, our estimate may be off by an order of magnitude in either direction.

On an individual level, many reasons for under-reporting are easy to understand because they involve motivations we might all have experienced. For example, one does not want to accuse falsely. One may also fear that reporting would take time away from research, or have concerns and fears about possible retaliation. One may assume someone else will or should report it. Or one may have sympathy towards a researcher, and might think "it's not too bad", it can be sorted out without a career-damaging investigation. Reporting also necessitates confidence that the issue will be examined carefully and thoroughly.

Keeping it quiet

The leaders of institutions may also have concerns about handling research misconduct. Because public image is important to institutions, some may try to minimize reporting and keep unfavourable information from reaching the ORI and the press. An institution may choose to ignore conducting an investigation and instead they may simply dismiss an accused person or even a whistleblower in the hope that the problem will go away without needing further examination. Additionally, institutional leaders may wish to ignore or minimize allegations of possible research misconduct to protect the revenue that the researcher generates; some may avoid investigations because they are costly in terms of time and money. Administrators may not recognize the significance of evaluating research misconduct and of course they may be poorly equipped to conduct an investigation in an appropriate manner.

Fundamentally all explanations seem to

share a common denominator — the failure to foster a culture of integrity. An analysis commissioned by the ORI found in 2000 that only 29% of institutional misconduct polices explicitly obligate members to report scientific misconduct[8]. Individuals and institutions, not the federal government, are the guardians of research integrity. Therefore, we urge action and recommend six strategies to champion integrity.

Adopt zero tolerance

To create a zero-tolerance culture, we think that it is essential that an institution specifies and implements the requirements that all suspected misconduct must be reported, and all allegations must be thoroughly and fairly investigated. Social responsibility to the academic community and to the public who fund the research will be strengthened when it is apparent that an institution has a real commitment to integrity.

Protect whistleblowers

Careful attention must be paid to the creation and dissemination of measures to protect whistleblowers. Responders to our survey said that reporting would be most likely to improve if institutions and the federal government increased the whistleblower protection. Indeed, more than two-thirds of whistleblowers, in a Research Triangle Institute study, experienced at least one negative outcome as a direct result of their actions[9]. Plus, 43% reported that institutions encouraged them to drop the allegation.

Clarify how to report

Researchers in our study also emphasized what would promote reporting: establishing a reporting system that clearly identifies the individuals to whom allegations should be brought, and establishing clear policies, procedures and guidelines related to misconduct and responsible conduct.

Train the mentors

If we want to build a stronger culture of integrity, then the current generation of researchers has to be educated to pay more attention to how they work with their junior team members. Social science has a long history of describing how group standards affect individual behaviour. Mentors specifically need to become more aware of their roles in establishing and maintaining research rules and minimizing opportunities to commit research misconduct[10]. Only 34% of scientists in a study with 2,206 laboratory directors strongly agreed that their mentor had prepared them to be a good mentor to others[11]. An institutional investment in building better mentors is an important vehicle to promoting research integrity.

SUSPECTED MISCONDUCT: 201 CASES OBSERVED BY 164 SCIENTISTS

	Number of cases
Type of misconduct	
Fabrication or falsification	120 (59.7%)
Plagiarism only	73 (36.3%)
Unknown	8 (4.0%)
Rank of those suspected*	
Professor or senior scientist	44 (21.9%)
Associate professor or scientist	28 (13.9%)
Assistant professor or scientist	34 (16.9%)
Postdoctoral fellow	50 (24.9%)
Graduate student	29 (14.4%)
Other (includes 1 unknown)	24 (11.9%)
How it was discovered	
Directly observed	23 (11.4%)
Observed products	53 (26.4%)
Told first, then observed	60 (29.9%)
Other direct evidence	30 (14.9%)
Other	30 (14.9%)
Don't recall	1 (0.5%)
No answer	4 (2.0%)
Was it reported?	
Yes, reported by responder	49 (24.4%)
Yes, reported by someone else	67 (33.3%)
No, not reported	75 (37.3%)
Don't know	5 (2.5%)
No answer	5 (2.5%)

* Eight cases identified more than one person involved in incident.

Use alternative mechanisms

Institutions must start to use other means to protect the integrity of their studies. The Institute of Medicine recommends that "Universities should not rely upon formal complaints of scientific misconduct as the sole source of monitoring the integrity and quality of the research conducted under their auspices. They need continuing mechanisms to review and evaluate the research and training environment of their institution."[12] Auditing research records would be one such means. Mechanisms of review are needed to reduce deficient record keeping, improper protection of human or animal subjects or the utilization of questionable research behaviour[13].

Model ethical behaviour

People imitate the behaviour of powerful role models. Institutions successfully stop cheating, for example, when they have leaders who communicate what is acceptable behaviour, encourage faculty members and staff to follow the policies, develop fair and appropriate procedures for handling misconduct cases, focus on ways to develop and promote ethical behaviour, and provide clear deterrents that are communicated[14].

Nearly one generation after the effort to reduce misconduct in science began, the responses by NIH scientists suggests that falsified and fabricated research records, publications, dissertations and grant applications are much more prevalent than has been suspected to date. Our study calls into question the effectiveness of self-regulation. We hope it will lead individuals and institutions to evaluate their commitment to research integrity. ■

Sandra L. Titus is director of intramural research, Office of Research Integrity, 1101 Wootton Parkway, Suite 750, Rockville, Maryland 20852, USA. James A. Wells is director of the Office of Research Policy, University of Wisconsin-Madison, 205 Bascom Hall, 500 Lincoln Drive, Madison, Wisconsin 53706, USA. Lawrence J. Rhoades is the retired former director of the Division of Education and Integrity, Office of Research Integrity, 1101 Wootton Parkway, Suite 750, Rockville, Maryland 20852, USA.

1. www.nap.edu/html/obas/
2. http://ori.dhhs.gov/research/intra/documents/Investigations1994-2003-2.pdf
3. Swazey, J., Anderson, M. & Lewis, K. *Am. Sci.* **81,** 542–553 (1993).
4. St James-Roberts, I. *New Sci.* **72,** 466–469 (1976).
5. Rankin, M. & Esteves, M. D. *Nurs. Res.* **46,** 270–276 (1997).
6. Department of Health and Human Services *Public Health Service Policies on Research Misconduct* 42 CFR 93 (2005).
7. Lederhendler, I. National Institutes of Health, personal communication.
8. http://ori.dhhs.gov/documents/institutional_policies.pdf
9. http://ori.hhs.gov/documents/consequences.pdf
10. Adams, D. & Pimple, K. D. *Account. Res.* **12,** 225–240 (2005).
11. http://ori.dhhs.gov/documents/research/intergity_measures_final_report_11_07_03.pdf
12. Institute of Medicine *The Responsible Conduct of Research in Health Sciences* (National Academies Press, Washington DC, 1989).
13. Martinson, B., Anderson, M. & DeVries, R. *Nature* **435,** 737–738 (2005).
14. McCabe, D. L., Trevino, L. K. & Butterfield, K. D. *Ethics Behav.* **11,** 219–232 (2001).

A more detailed report discussing this study can be found at http://tinyurl.com/3keo6h.

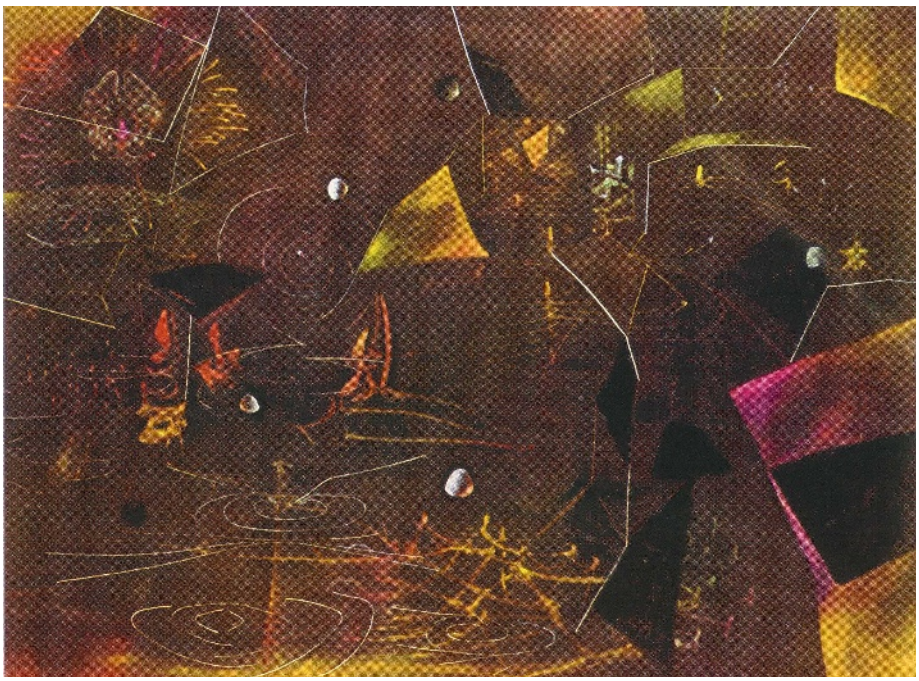
See Editorial, page 957.

BOOKS & ARTS

Quantum weirdness and surrealism

A joint exploration of early modern physics and the surreal art movement shows these twentieth-century revolutions had more in common than we thought, explains **Philip Ball**.

Surrealism, Art and Modern Science
by Gavin Parkinson
Yale University Press: 2008. 294 pp.
$60.00



Roberto Matta's *The Vertigo of Eros* conveys the collisions and confusions of modern physics.

Surrealist artists working in the early twentieth century, including André Breton, Max Ernst, Man Ray and Salvador Dalí, disorientated their audiences using odd, ambiguous juxtapositions and distortions of objects and images. Around the same time, relativity and quantum theory unsettled scientists with notions of plastic time and space, multiple truths and challenges to causality.

Surrealism, Art and Modern Science shows the links are explicit, not a superficial analogy applied in retrospect. The Surrealist artists referred repeatedly to relativity and quantum mechanics in their writings. Dalí intended his drooping watches to allude to distorted space-time, describing them as a "soft Camembert of time and space". Gavin Parkinson's study removes any contention about the connection between surrealism and physics, and makes you wonder why it was not made before.

An art historian at the Courtauld Institute in London, Parkinson paints a engrossing picture of the period between the 1920s and 1940s, when modernism flourished and created an intellectual ferment that spawned numerous highbrow journals wherein art and science met. He challenges the simple view of physicist and writer C. P. Snow that the 'Two Cultures' of art and science have nothing to say to each other, and, even if they did, no mutual language in which to speak it.

An important bridge between the Surrealists and physicists was *The New Scientific Spirit,* a book published in 1934 by the French philosopher Gaston Bachelard. Trained in chemistry and physics, Bachelard became steeped in the proto-Surrealist poetry of Paul Éluard and Paul Valéry. Seeing the progression of science as a series of jumps rather than gradual advance, a position now attributed to science historian Thomas Kuhn, Bachelard considered his own times to be a period of rapid intellectual change. During revolutionary epochs, he said, there is a moment when old notions have shattered but new ones have not yet crystallized, when one must consider seemingly irrational ideas that are propelled by their own momentum.

It is no wonder the new physics appealed to the diverse, volatile band that constituted the Surrealists. Motivated by deep, even ponderous, philosophical ideas, these artists wanted to discard the comfortable certainties of representational work while avoiding the retreat into mysticism that total abstraction threatened. For many Surrealists, art was a means of investigating the world, particularly the self. The influence of Sigmund Freud's psychoanalysis on them has been well documented, but physics offered vindication of their concepts too. The painter Roberto Matta said that Albert Einstein was as important as Freud to the modern artist, and filled his pictures with geodesic space-time grids like those now used to illustrate black holes and wormholes. The polymathic Valéry, Breton's mentor, cultivated friendships with physicists Paul Langevin and Louis de Broglie, Jean Perrin, Niels Bohr and Einstein.

Physicists sometimes reciprocated. Marie-Antoinette Tonnelat, de Broglie's colleague, said in 1952 that "In physics as in painting, Surrealism denies the possibility of a description which does not carry explicitly the stamp of the observer" — we construct what we see. In 1934, philosopher Henri-Charles Puech wrote that in modern physics "the exact delineation of reality gives way to a vaguer consideration of unities more or less arbitrarily defined."

No radical artist could resist this message. Thus, Surrealists sought to enlist physics to unbalance our preconceptions. If you think our pictures are absurd and bewildering, they said, just look at what the physicists are saying. Some artists betrayed a kind of science envy: Dalí spoke of science's "burning analytic precisions". Both he and Breton tried to steer a path between art and science, maintaining ambiguity about their true goal.

Were these appropriations of science merely misconceived analogies? Parkinson, whose introduction to the physics of this era is splendid and nuanced, is aware of that danger, as were some of the artists. The Viennese painter and writer Wolfgang Paalen criticized some fellow Surrealists for using science simply as poetic ornament. Other artists cloaked their superficial understanding in the opaque verbal blanket for which French philosophy later became notorious. Breton's partial grasp of physics did nothing to check his arrogant appropriation of it: after he interpreted Einstein as saying that "one event can be the cause of another only if they can both be brought about in the same point in space", he blithely added "that is what I have always thought."

The Surrealists' interest in physics was genuine. But *Surrealism, Art and Modern Science* gives the impression that it was one of several themes commandeered for, and then shoehorned into, a radical social and political agenda. Breton and his fickle coterie flitted between Werner Heisenberg's uncertainty principle, Marxism, magic and occultism. This need not be problematic if they were simply looking for artistic inspiration, but Breton's intent was to make statements about the nature of reality.

One difficulty is that the scientific dilettante often converts particulars to generalities. What applies under one special, constrained set of circumstances is held up as a principle applicable to all things. Relativistic distortions and quantum indeterminism become universal attributes. Dalí, for example, spoke of a "psychic dilation of ideas" — as if our adherence to classical concepts were a conservative, bourgeois delusion rather than a necessary approximation. Parkinson describes how, when physicists such as Arthur Eddington used everyday analogies for pedagogy, their artist readers took them literally.

Perhaps we should not listen to what the artists say, but look at what they do. In his painting *The Vertigo of Eros*, Matta conveys as well as anything I have seen the collisions and confusions of the new sciences, combining multiple reference frames with allusions to Hermann Minkowski's bent space-time and to particle physics. The picture does not precisely illustrate, still less illuminate, the science that inspired it. It creates a nexus of reference points that sets the mental pathways buzzing. That is surely what good art does. ■

Philip Ball is a consultant editor for *Nature*. His latest book is *Universe of Stone*.

Wish you were here?

A Nuclear Family Vacation: Travels in the World of Atomic Weaponry
by Nathan Hodge and Sharon Weinberger
Bloomsbury: 2008. 336 pp. $24.99, £12.99

How are you spending your next holiday? Tired of the same old thing? You might want to pick a different destination from *A Nuclear Family Vacation*, a new book and travel guide by veteran defence reporters Nathan Hodge and Sharon Weinberger.

This husband-and-wife team take the reader on a rapid, darkly comic tour of nuclear weapons sites across the world. A rare achievement in a nuclear policy book, their narrative demystifies an intimidating topic for a broad audience without sacrificing substance.

Instead of pontificating on thermonuclear war, Hodge and Weinberger give us an eye-level view, often through their car window. They take us to former Soviet testing grounds in Kazakhstan, missile defence sites on remote Pacific islands and nuclear laboratories around the United States, including once-secret nuclear bunkers built to shelter dignitaries in the Catoctin Mountains, 100 kilometres north of Washington DC. The couple meets the people who work there and listen to their stories.

US Air Force missile men still work three-day shifts in underground silos, ready to launch nuclear warheads with 15 minutes' notice. Workers at the Y-12 complex in Oak Ridge, Tennessee, are tearing down the buildings in which cold-war arsenals were assembled, and constructing new ones for another generation of warheads. Scientists at labs nationwide continue to dream up new weapons, even with the lack of a clear and present danger. Gerold Yonas of Sandia Laboratory in Albuquerque, New Mexico, was one of the architects of the Strategic Defense Initiative, or 'Star Wars programme', a proposal to protect the United States and its allies from attack by nuclear-armed missiles. He admits that his career "has not been marred by a single success", but past design failures have not stopped him from dreaming up new doomsday weapons. The book paints a powerful portrait of a sprawling, decaying nuclear complex struggling to perpetuate itself without a clear purpose.

Iran's Isfahan atomic facility : a hot destination for nuclear tourists.

H. FAHIMI/AFP/GETTY IMAGES

When *A Nuclear Family Vacation* is good, it is very good. Hodge and Weinberger tackle the suitcase nuke, a small nuclear weapon that could be delivered by a single soldier, something I get asked about at nearly every lecture. They describe the closest US attempt to build such a device, the Special Atomic Demolition Munition, by taking us to the National Atomic Museum in Albuquerque, where a mock-up of the weapon is on display. They discuss the history of other tactical nuclear weapons. My favourite is the Davy Crockett, a nuclear bazooka designed to fire a sub-kiloton warhead 2 kilometres — the US Army eventually realized this was not such a good idea. These stories bring the issue alive.

Hodge and Weinberger interview scientists, bureaucrats and politicians to flesh out daily life in the nuclear weapons complex, including the labs at Los Alamos and Sandia in New Mexico, Lawrence Livermore in California, and Oak Ridge. Standing alone in a vast desert, settled alongside a major progressive city, or tucked into the Appalachian mountains, these remnants of cold-war infrastructure battle to define their new role. Once bustling centres of activity are skeletons of their former selves. Safety and security standards have dropped and gifted scientists are drifting away.

But the laboratories and weapons are not fading away completely: the money is still flowing. For many communities, the labs are an important source of jobs. For many officials, nuclear capability is still central to the national security strategy. The United States spends more than $54 billion each year on nuclear weapons and related activities. Plans for a reliable replacement warhead, or RRW, would enlarge the weapons production facilities. The new warhead is the cornerstone of an ambitious expansion plan called Complex 2030, in which the United States would ramp up its ability to produce nuclear weapons and design and field thousands of new warheads. At a price. William Hartung of the New America Foundation think-tank conservatively estimates that "the full costs of Complex 2030 could easily reach $300 billion ... a $125-billion increase over the estimated costs of maintaining the current weapons complex."

Future plans must take into account more than fluctuating budgets. Hans Kristensen

of the Federation of American Scientists believes that "the RRW and Complex 2030 programmes are not only unnecessary. They also undercut efforts to convince non-nuclear nations to forgo nuclear weapons and to convince new weapons states such as India and Pakistan to refrain from developing additional warheads."

The concept of nuclear deterrence is dead, according to former US commander-in-chief of the Strategic Command, General James Cartwright, now Vice Chairman of the Joint Chiefs of Staff. In place of mutually assured destruction, Hodge and Weinberger describe strategic deterrence, an "amorphous concept" that covers everything from "a public relations campaign to a bunker buster". The authors puzzle over why officials still cannot provide answers to the basic questions of how many nuclear weapons are wanted, what they will be used against, and when.

An equally puzzled Congress is stalling on the issue. It has cut the budget for RRW and the Complex 2030 expansion in response to the concerns of scientists and citizens, but the projects are still alive. Managing the programmes and their effects on global nonproliferation efforts will be one of the first items on the next president's nuclear agenda.

Travelling beyond the vast weapons complex of the United States, the authors visit the countries you would expect, and some you would not. The former Soviet nuclear cities, mirrors of Oak Ridge and Los Alamos, are even more decayed and dilapidated. Security analysts have been troubled by unsecured facilities and underemployed weapons scientists since the fall of the Soviet Union. Joint programmes between the United States and Russia have secured around half of the poorly protected weapons and materials in the former Soviet states, but much material is still vulnerable to terrorist theft. It also leaves unaddressed a legacy of nuclear pollution. Kazakhstan, the authors discover, gave up its deployed nuclear weapons but still suffers from nuclear isotopes in its soil.

Hodge and Weinberger succeeded in entering Iran, whose nuclear energy programme could bring it a weapons capability. The Iranian guides leading a tour of the Isfahan Uranium Conversion Facility, which I visited in 2005, tried to convince the authors that the programme was peaceful and transparent. "Our trip," they conclude, "could be taken as a sign of openness, but it could just as easily be interpreted as cynical propaganda." They give a nuanced understanding of the programme from the Iranian point of view — both as a source of national pride and a burden that keeps the country sanctioned and isolated. Iran, like other nations, has yet to come to terms with the nuclear programme it has created.

A Nuclear Family Vacation has its flaws. The opening chapters drag, and the discussion sometimes meanders, determined by the material the authors were able to get rather than what is most important. Overall, the book sparkles with anecdotes and insights. It is well worth the trip. ■

Joseph Cirincione is the president of the Ploughshares Fund and author of *Bomb Scare: The History and Future of Nuclear Weapons*.

Memories revisited

In Search of Memory: The Neuroscientist Eric Kandel
Film directed by Petra Seeger
Showing in Austria.


P. SEEGER

German filmmaker Petra Seeger met neuroscientist Eric Kandel by chance in Berlin two years ago, and was enthralled by his research and life story. An Austrian Jew forced to flee the Nazis in 1939, Kandel (pictured) is still coming to terms with his traumatic past. *In Search of Memory,* Seeger's 95-minute documentary of the mischievous 79 year old, premiered on 26 May in Vienna, Kandel's childhood home.

In 2000, Kandel shared the Nobel Prize for Physiology or Medicine for his work on how neurons lay down memories. The film weaves Kandel's recollections and the science of learning and memory. Seeger accompanies him to Vienna to seek out his family's old apartment, his father's toy shop and other poignant places that he has avoided for fear of stirring up pain. Seeger's camera follows Kandel in mid close-up. Re-enactments of his childhood are mixed with archive footage from 1930s Austria and contemporary scenes shot in his laboratory at Columbia University in New York.

Kandel recalls on screen the jubilant welcome of Hitler's troops as they marched into Vienna in March 1938. The following November, a few days after his ninth birthday, he witnessed the horrors of Kristallnacht, when rioters destroyed synagogues and Jewish premises. Instructed to leave their house, his family returned two weeks later to find that everything had been stripped from it, even his birthday toys. He recalls the collusion of many Austrians with the Nazis and the lack of support, or even sympathy, from former non-Jewish friends.

Those enduring memories fuelled Kandel's desire to understand the biological basis of memory. His approach was unfashionably reductionist. He chose as his model organism the sea slug *Aplysia californica*, which has just a few large nerve cells and a robust reflex — it withdraws its gills in response to stimulation. Kandel showed that the sea slug can learn to modify this reflex when repeatedly stimulated, and that the change is caused by strengthening of the synapses, the regions where neurons connect with each other.

Seeger mirrors that reductionist approach in her filming. She eschews high-tech animations, relying on Kandel to explain his science with only a flip chart and a large plastic model of the brain. His lab seems busy and his colleagues look happy: young researchers relate their own discoveries with moving enthusiasm. Scientists come across as vibrant people with pasts, sensitivities and futures — with stories to tell.

The naive viewer may not take home many scientific details, but the documentary conveys the breadth of neuroscience and the scientific process. It describes different types of memory that arise in distinct parts of the brain, and the fundamental cellular process of synapse strengthening. Kandel's story shows that to reach the truth, you sometimes need to go back to basics before reconstructing the big picture.

Kandel laughs a lot as he confronts his past. In one memorable sequence in New York, he asks an old man seated on a chair in the street if he remembers his father's store. "What's your problem?" the old man barks. But within a minute they banter about their age, and laugh so infectiously that the audience laughs too. In another scene, Kandel's eyes brim with tears.

The strength and weakness of *In Search of Memory,* named after Kandel's book of the same name, is that it is uncritical, a love affair with this petite, demanding genius. The film manipulates its viewers into adoration. Scientists outside Kandel's lab barely get a mention. But the warmth of the film, together with the political and scientific importance of the subject, more than compensate. ■

Alison Abbott is *Nature*'s Senior European Correspondent.

Building nations after conflict

Fixing Failed States: A Framework for Rebuilding a Fractured World
by Ashraf Ghani and Clare Lockhart
Oxford University Press: 2008.
264 pp. $24.95, £14.99

Political instability in collapsed or collapsing states is one of the greatest sources of human misery. Since the end of the cold war, within-state conflicts, such as civil wars and separatist rebellions, have caused ten times more deaths than have wars between states. The indirect, non-military consequences of internal conflict — civilian casualties, refugees, wrecked economies, famine and disease — are orders of magnitude worse than the direct, military outcomes. In *Fixing Failed States*, former Afghan finance minister Ashraf Ghani and development-policy expert Clare Lockhart analyse why states crumble and propose a framework for rebuilding state capability.

International organizations, such as the United Nations, International Monetary Fund and World Bank, have been busy recently. Before 1990, the United Nations initiated a peace-keeping mission on average once every three years. Since 1990, they have run approximately three new missions each year. Many interventions successfully stopped active fighting. Yet our record in addressing the deep causes of conflict and constructing viable post-conflict societies has been poor.

Ghani and Lockhart are optimistic that we have "the elements of a new approach to state building". They review four examples — post-war Europe, Singapore, the southern United States and Ireland — that, in their opinion, prove that countries confronted with devastation, chaos and entrenched poverty can transform themselves into prosperous and stable members of the global community. Apart from Singapore, however, these are not examples of state collapse. Europe in 1945 was devastated by interstate war; Ireland was poor before its economic miracle but not a collapsed state; and few would consider the United States to be weak. Seceding from the Federation of Malaysia in 1965, Singapore was plagued by poverty, corruption and a communist insurrection. Today, it is an economic powerhouse with one of the least corrupt civil services in the world.

A better example, dealt with in passing, is the remarkable reversal of China's fortunes during the twentieth century. We think of China now as a surging world power that has enjoyed 10% annual economic growth since 1980. Yet in the 1920s and 1930s, China was a collapsed state. With a weak national government and most of its territory controlled by warlords, it was affected by incursions from Europe and Japan, and by a raging communist uprising.

Can general lessons be drawn? Ghani and Lockhart fail to present a compelling analysis, quoting instead political leaders ranging from UK prime ministers William Gladstone and Gordon Brown to Mahathir Mohamad, former prime minister of Malaysia.

The authors do not use knowledge accumulated by social scientists in recent decades. There is no mention of the burgeoning literature on revolutions and state collapse, such as the work of sociologist Jack Goldstone, or of books on transitions from civil war to democracy, such as Elisabeth Jean Wood's *Forging Democracy from Below* (Cambridge University Press, 2000).

Such recent studies indicate that generalities can be found in the historical record, although we are only beginning to understand them. One generality that might have been included is the role of political élites in state building. When élites are fragmented — perhaps by ethnic divisions or rival patronage networks — each faction focuses on getting its own piece of the common pie instead of increasing the total size of the pie. Conflict among élites prevents consistent government policy and often mutates into an armed struggle. A necessary condition of a strong, effective and democratic state is some degree of consensus among the élites about the strategic goals and the political process for achieving them.

One of the most important social mechanisms that determines whether élites are consolidated or fragmented is élite overproduction — growing numbers of aspirants for élite positions, resulting from demographic expansion and unbalanced social mobility. When the number of aspirants greatly exceeds the supply of élite positions, fragmentation and conflict ensue.

Another important issue is cooperation and trust. A dysfunctional state is like a vast game of prisoner's dilemma in which those who cooperate are taken advantage of, leaving withdrawal from cooperation as the only rational strategy. Citizens avoid paying taxes because this will only enrich corrupt officials, whereas officials spend public money on themselves and their clans because if they do not, others will take it. How can societies escape such a state of pervasive distrust? We lack a good theory to provide an answer, but there are promises of one.

The new discipline of experimental economics demonstrates how non-cooperative groups

Winning Darwin design takes root

Sculptor Tania Kovats has won the Darwin's Canopy competition to design a new permanent artwork for Charles Darwin's bicentenary at the Natural History Museum in London. A longitudinal cross section of a real fallen oak tree, including roots, trunk and branches, will be veneered into the panelled ceiling of a gallery behind the central hall.

Ten contemporary artists submitted designs, which are on display at the museum until 14 September. Kovats's ceiling installation will be unveiled on 12 February 2009, on what would have been Darwin's two-hundredth birthday.

Kovats, currently tracing Darwin's footsteps in South America, has long marvelled at a piece of petrified tree held at the museum. A branching sketch in Darwin's notebook of 1837 also

NAT. HIST. MUS., LONDON

inspired her. "Whether a tree or a coral," she notes, it is "quite remarkable for how it represented to him a proof of where his thoughts were going".

Within Kovats's design (sketch, pictured), her tree's roots will represent the research of museum scientists. Its branches will represent the museum's role in disseminating knowledge, as well as taxonomy, and the use of a fallen tree hints at how Darwin felled the existing orthodoxy. The tree "is a real thing as well as a sculptural intervention, and as such can take its place amongst the other real things housed in the collection", she explains. ■

Colin Martin is a writer based in London.

can shift to being cooperative by applying moralistic punishment, such as sanctions, against defectors. On a national scale, history suggests that external pressure applied to a society may increase internal cohesion and cooperation. National humiliation of China, first from the European great powers in the nineteenth century and then from Japanese occupation during the Second World War, played an important part in its post-war reunification, for example.

The policy implications of historical outcomes are doubtful. We can hardly subject societies to horrific stresses deliberately, and they may produce oppressive regimes in response. Rather than focus on a few haphazard cases, systematic research is needed to find out what works. Although not mentioned in *Failed States*, such programmes are currently being conducted by, for example, the Political Instability Task Force in the United States and the Centre for the Study of Civil War in Oslo, Norway.

Ghani and Lockhart propose an agenda for state building, but their weak analysis undermines its credibility. They suggest a 'sovereignty strategy' that involves formulating a strategy, then setting the goals and rules of the game, mobilizing resources, allocating critical tasks and, finally, monitoring implementation of the strategy. This generic approach does not suggest concrete policies. For example, the book describes how a strategy formulated in the Indian state of Andhra Pradesh "forced a sobering reading of conditions: corruption, inefficient use of state resources, short-term planning and poor infrastructure. This reading of context enabled participants to embrace change and leaders to set a clear sense of direction." Given such an easy buy-in, one wonders why this approach has not enabled more sides, such as the Maronite Christians and the Shia and Sunni Muslims in Lebanon, to make peace given the many opportunities they have had to "embrace change".

I nonetheless commend Ghani and Lockhart for raising this issue. We cannot afford to ignore failed states. We insist that new drugs are exhaustively tested before they are used, so shouldn't we invest in better social science before we intervene with failed states? Otherwise, our well-intentioned but misguided attempts to fix them may be as helpful as the medieval practice of blood-letting. ■

Peter Turchin is professor at the Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043, and author of *War and Peace and War*.

J. KOHEN/WIREIMAGE

Q&A: Insight into Einstein

Actor **Alan Alda**, who starred in the television series *M*A*S*H* and now hosts *Scientific American Frontiers* on US network PBS, is fascinated with physics. At last month's World Science Festival in New York he led a panel discussing the quantum world, portrayed Richard Feynman in the play *QED*, and presented *Dear Albert*, his new play drawn from Albert Einstein's letters.

Why did Einstein's letters interest you?
It's very important for us to see that science is done by people, not just brains but whole human beings, and sometimes at great cost. Letters can be very personal, and sometimes confrontational.

I had also planned to write a play about Marie Curie's letters. I got a little discouraged because not only are they in Polish and French, but the French letters are still slightly radioactive. After you look at them they go over you with a Geiger counter. I thought I'd wait until somebody else goes in a hazmat suit and translates them. So I stuck with Einstein.

Einstein emerges from your play as a highly volatile character, sometimes spiteful and domineering, sometimes withdrawn and resigned. How do you see him?
Einstein claims not to have felt lonely, but he was a lonesome figure. He could see far out into the cosmos but he was myopic about the people next to him. It was difficult for him to take the time for what he called the "merely personal". And he really did seem to take refuge in these very complicated images in his head. Like Feynman, he challenged every idea that came to him. He wanted to rethink it, he wanted to see more deeply into it.

Why did you focus on Einstein's relationships with his two wives, Mileva and Elsa?
Plenty of his correspondence with colleagues was about the science that he was working so hard on. But I wanted to show the personal side of the discoveries and ruminations. For somebody with hair like that, he did awfully well with the women. At one point he couldn't decide whether to marry his second wife Elsa or her daughter Ilse, who wrote to a friend, "Albert refuses to take a position on this".

Will the play be performed again?
I don't know. It was like a high-energy experiment: we just let the actors collide with the material. Whatever particles came out of it we could observe for a short time, and now it has evaporated. ■

Interview by **Jascha Hoffman**, a writer based in New York.

ESSAY

Playing by numbers

Statistical analysis can inform the history of music, classification technologies, and our understanding of the act of composition itself, argues **Damián Zanette**.

Music history is riddled with debates on attribution. Did Andrea Luchesi compose many of the symphonies currently attributed to Mozart? Is the score of *L'Incoronazione di Poppea* Monteverdi's, or the collaborative work of several editors during its early performances across Italy? Did Johann Sebastian Bach really write the chorale *Nun ist das Heil und die Kraft*, the original score of which has never been found?

Statistical analysis may help to resolve such long-standing controversies, as it has proved successful in linguistic texts. It may also allow electronic databases to automatically classify musical style and period. And it promises much more — to help us understand some of the most elusive qualities of music, their connection to its organizational structure and to the cognitive processes involved in both the composition and perception of music. Eventually, statistics may also allow us to identify a quantifiable signature of complexity in music.

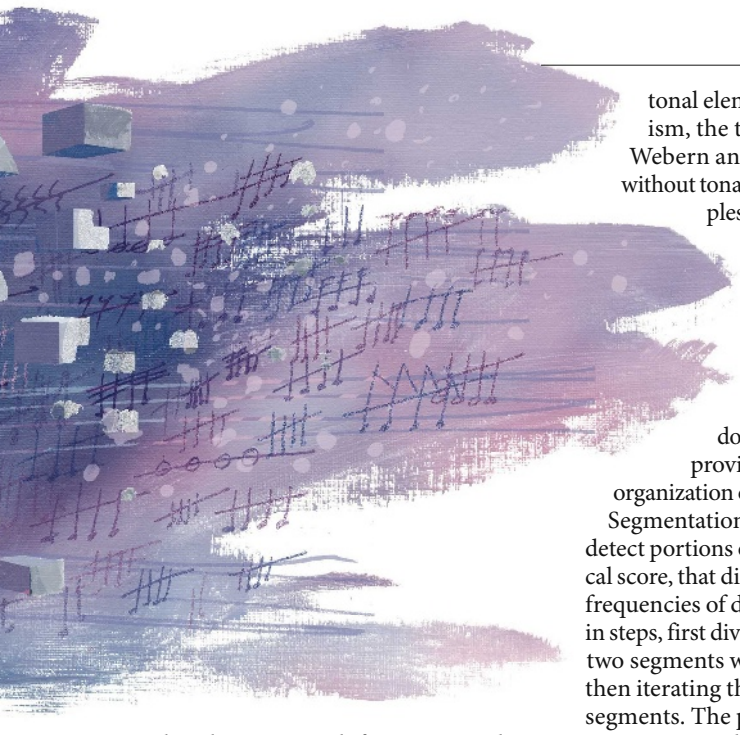
The composer Arnold Schoenberg summarized the fundamental principles of musical form as "the demand for repetition of pleasant stimuli, and the opposing desire for variety, for change". Repetition of melodic motifs, rhythmic patterns and harmonic progressions makes musical structure coherent and forms the basis of its comprehensibility. Variation, in turn, keeps monotony and dullness at bay.

This delicate balance — somewhere between the uniform ticking of a clock and the random pitter-patter of raindrops — is reminiscent of complex systems, in which an intermediate degree of internal organization maintains coherence, yet allows for rich dynamics and functional flexibility. The melodic and rhythmic patterns of even the simplest folk tune reveal the complexity of the creative process, and of the system behind it — the human brain. By probing the structural texture of music and the recurrence and diversity of elements such as notes, rhythms, melodies and chords, statistical techniques provide a way to penetrate the nature of mind.

Word play

In the 1930s, the American philologist George Zipf discovered a strong regularity in the relative frequencies of word occurrence in speeches and texts. Now called Zipf's law, this rule applies to many different authors, styles and languages. If, for instance, the tenth most used word in a text occurs 300 times, Zipf's law predicts that the hundredth most used word will appear some 30 times.

In 1955, social scientist Herbert Simon pointed out that Zipf's law can be quantitatively explained by assuming that the usage frequency of a word increases proportionally to its previous appearances — the more you use a word, the more you will use it. This very



D. PARKINS

simple rule was enough for Simon to derive Zipf's law as the inverse relation between the number of occurrences of a word and its rank in frequency of use.

Since the late 1980s, several researchers have shown Zipf's law also holds for musical elements within pieces. Words are replaced by notes, defined by pitch and duration, or by composite items such as note duplets and triplets, interval successions and chords. This suggests a strong affinity between the processes of writing text and composing music.

Simon's model for the relative frequency of words in a text can be interpreted as representing the progression of the author's choices during the creative process that shape the work's intelligibility. In language these choices are grammatical, morphological and semantic. In music they are melodic, harmonic, rhythmic and dynamic.

Music as message

Literary texts and musical compositions are created as organic entities, not series of isolated decisions. Nevertheless, the outcome is an ordered sequence of events conveying information: a message. As the message flows, a context emerges, favouring the appearance of some elements at the expense of others. From this viewpoint, Simon's model for Zipf's law unifies the concept of context in both language and music.

We can also distinguish the choices made by composers from the different forms that Zipf's law takes in their music. The law quantifies the difference, say, between the intentional lack of tonal context in Schoenberg's pieces, and Bach or Mozart's more consistent, less flexible use of tonal elements. Yet intriguingly, serialism, the technique Schoenberg, Berg, Webern and others used to write music without tonality, is still based on the principles of repetition and variation.

Zipf's law is not the be-all and end-all of the statistical characterization of musical structure — for one thing, it would still hold if all the notes of a composition were shuffled and rearranged at random. Happily, information theory provides other ways to analyse the organization of symbols in a sequence.

Segmentation, for instance, can be used to detect portions of a sequence, such as a musical score, that differ as much as possible in the frequencies of different symbols. It proceeds in steps, first dividing the whole sequence into two segments with maximal difference, and then iterating the algorithm on the resulting segments. The product is a dissection of the sequence into domains which are maximally divergent — the relative frequency of symbols differs as much as possible between the resulting domains.

Segmentation was recently used to analyse the first movement of Mozart's keyboard sonata in C major (K. 545), as a sequence formed by the twelve tones of the chromatic scale. The analysis revealed the same tonality changes spotted by humans trained in musical analysis. In 1997, psychologist Carol Krumhansl of Cornell University in Ithaca, New York, demonstrated that non-specialist listeners also spot modulation between different tonalities when asked to divide Mozart's keyboard sonata in E-flat major (K. 282) into sections with different perceived musical qualities.

> **"Statistical techniques provide a way to penetrate the nature of mind."**

These preliminary results suggest that such statistical tools, which can be automated for large-scale computational application, can reveal the same structural features as ordinary methods of musical analysis. They might also unveil evidence of hitherto undetected organizational levels and patterns.

Segmentation can also be applied to combinations of pitch and duration, dynamics, intervals and chords. Analysing these more complex items may reveal patterns related to the richer cognitive qualities of music, such as melodic inflections and rhythmic change, which listeners associate with the unfolding of a piece's mood. This has its limits. As a collection of symbols becomes larger and more sophisticated, each element's frequency decreases. When each symbol appears too few times to be statistically significant, a meaningful message becomes indistinguishable from a random sequence.

Complex futures

Since 1996, physicist Pedro Bernaola-Galván of the University of Málaga, Spain, and his collaborators have applied segmentation analysis to DNA sequences to study the origin and significance of long-range nucleotide patterns. The resulting segments show large variations in length, a feature that has been related to the slowly decaying probability that two nuleotides of the same type are found at a certain distance in the genetic sequence. Galván's group has suggested that a broad distribution of segment lengths may be a signature of complexity for symbolic sequences. Both random and periodic sequences, such as raindrops' pattering and clocks' ticking, show little variation in segment lengths. The information-carrying DNA sequence, on the other hand, is characterized by a long-tailed distribution.

Mathematicians should investigate whether segmentation of long linguistic and musical sequences also gives broad length distributions. This would enable us to compare the degrees of complexity in language, music and the genetic code, disclosing structural similarities and differences between these forms of communication. A quantification of complexity in music would also allow us to identify the structural elements underlying different periods and styles. But for the time being, the quest to define a unified complexity measure continues.

Statistical analysis seems to be at odds with traditional ways of thinking about art. These — unlike mathematics — emphasize aesthetic nuances, psychological and experiential qualities and personal values. Indeed, how we integrate and elaborate sensory information into artistic experience may always be beyond quantitative description. Nonetheless, quantitative methods can tell us much about artistic creation — notably, about the organization of an artwork's many strands into a comprehensible structure. ■

Damián Zanette is head of the Statistical and Interdisciplinary Physics Group of Centro Atómico Bariloche and professor of physics at Instituto Balseiro, Avenida Ezequiel Bustillo 9500, 8400 San Carlos de Bariloche, Río Negro, Argentina. He is a co-author of *Emergence of Dynamical Order*.

For further reading see http://tinyurl.com/3k88hs. See other essays in the Science & Music series at www.nature.com/nature/focus/scienceandmusic.

ESSAY

A century of puzzling

Believed to be the world's first printed document, the Phaistos Disc was unearthed 100 years ago. Andrew Robinson explains why this remarkable object remains undeciphered.

The Rosetta Stone is the most famous of ancient inscriptions; it unlocked the meaning of thousands of Egyptian hieroglyphic inscriptions. The undeciphered Phaistos Disc, discovered by an Italian archaeologist at Phaistos near the coast of southern Crete a century ago next month, is perhaps the most infamous.

Luigi Pernier found the disc on 3 July 1908 in a basement cell of a ruined Minoan palace dating from the first half of the second millennium BC. No other samples of the script have turned up since. Pernier published his find in 1909 without trying to decipher it. The same year, archaeologist Arthur Evans, discoverer of ancient Knossos in the island's north, included fine photographs and good drawings of the disc in an appendix to his pioneering volume of Minoan inscriptions, *Scripta Minoa*. Tantalized scholars in many countries began to speculate as to its meaning. Evans's subsequent lengthy discussion of the disc in volume one of his celebrated *The Palace of Minos at Knossos*, published in 1921, threw down the gauntlet to would-be decipherers everywhere.

Over the past hundred years the disc has become notorious for three reasons. First, the pictograms on its clay surface have provoked dozens of wildly incompatible hypotheses about what it is and what it says. Competing interpreters have included a Cambridge classicist, a Harvard professor of zoology and, very recently, a geneticist from the University of Perugia in Italy. Interpretations range from astronomical calendars and bronze-age computers through board games to a victory chant and pre-homeric poetry, written in languages as disparate as Greek, Minoan, Hittite, Semitic, Egyptian and Slavonic. In the 1980s, the classicist John Chadwick, who helped to decipher the Minoan script Linear B, received roughly one claimed decipherment per month.

Around the same time, *National Geographic* even planned a lead story supporting a decipherment of the disc as a Minoan proclamation of war against Anatolia written in a 'Hellenic' dialect. Three senior classicists led by Chadwick persuaded the magazine to withdraw its embarrassing endorsement. Another of the three, Emmett Bennett Jr of the University of Cincinnati, wrote in 1998 that any book cover emblazoned with the Phaistos Disc — and there have been many — was for him "the equivalent of the skull and crossbones on the bottle of poison". In 2000, the *American Journal of Archaeology* ran a review titled "How not to decipher the Phaistos Disc", by Yves Duhoux of the Catholic University of Louvain in Belgium, author of the leading scholarly book on the subject, *Le Disque de Phaestos*.

Second, the disc is notorious for being the world's first 'printed' document, predating Gutenberg's Bible by more than 3,000 years. As Jared Diamond explains in *Guns, Germs and Steel*, this is "a threatening challenge to historians", because it suggests that the history of invention is so idiosyncratic as to be unpredictable. How could printing, once invented, disappear for millennia?

Third, the disc is a Greek national icon, and a key attraction at the Archaeological Museum at Iraklion in Crete. The Greek authorities have rebuffed several appeals, most recently in 2007, for the disc to be thermoluminescence tested — a technique that reveals an approximate date of last firing for pottery. The reason, says Jerome Eisenberg, an expert in ancient forgery and fraud, and editor-in-chief of the international art and archaeology review *Minerva*, is that "no Greek scholar or

G. DAGLI ORTI/CORBIS

politician would dare to help 'destroy' such a national treasure".

The archaeological context of the disc's discovery implies a date of 1850–1600 BC. To suggest that it might be a 1908 hoax — Piltdown man with a printing set — as two or three scholars, including Eisenberg, have proposed, is as heretical to Greek ears as the international scholarly allegations of the 1990s that Heinrich Schliemann may have faked some discoveries at Troy and Mycenae.

Spiral stamps

The disc is made of fine clay. It is about 16 cm across and 1.9 cm thick. Both sides carry an inscription arranged in a spiral around the centre — characters impressed with a punch or stamp before the clay was fired. There are 241 or 242 characters (one is damaged), which comprise 45 signs of variable frequency. For comparison, there are thousands of characters in a few pages of printed English text, comprising the 26 signs we call letters. Lines partition the disc's characters into 31 short sections on side A and 30 on side B, most of which contain three, four or five characters. It is tempting to speculate that these sections represent words in the language of the disc.

That the characters were printed, not carved, is beyond dispute. But no one knows why the disc's maker bothered to produce a punch or stamp for each sign, rather than inscribing each character afresh. Egyptian hieroglyphs or Mesopotamian cuneiform of the second millennium BC are inscribed on stone or clay; ditto the Minoan scripts Linear A and B found at Phaistos, Knossos and other Cretan sites. If the punch or stamp was to 'print' many copies of documents, one would expect further samples to have turned up in a century of intensive Mediterranean excavation.

There is patchy and inconclusive evidence for and against the disc's Cretan origin. The signs look nothing like those of Linear A, Linear B or any other Minoan script, except coincidentally. This has led some, including Evans and Chadwick, to propose that the disc — and presumably its language, too — was an import. One sign bears a remarkable resemblance to the architecture of rock tombs found in Anatolia in modern Turkey. One or two others resemble signs found on a few contemporaneous objects from different sites in Crete. Most scholars today, including Duhoux, think it a plausible working hypothesis that the disc was made in Crete.

The puzzling artefact was almost certainly written from the rim to the centre. The impressions show that in some cases a character very slightly overlaps that to its right. This must mean that the scribe wrote the characters from right to left, probably revolving the disc for convenience. And a right-to-left direction is feasible, given the order of the characters, only if the disc was inscribed from rim to centre. Presumably, it was meant to be read in the same direction.

ART ARCHIVE/BIBLIOTHÈQUE DES ARTS DÉCORATIFS PARIS/G. DAGLI ORTI

Archaeologist Arthur Evans tantalized scholars with his description of the Phaistos Disc.

The character count and the number of signs tell experienced cryptographers two things (assuming that the writing represents a spoken language, rather than specialized notation as in a calendar or a game). First, the low ratio of character count to number of signs — compare the higher ratio in even one page of printed English — means we do not have enough text to decode it without help from other clues, such as archaeological context or knowledge of the likely underlying language. Computers are of no help here as they depend on statistical analysis of ample text.

> **"To make further progress on the Phaistos Disc we must hunt for more examples around the shores of the eastern Mediterranean."**

Second, more helpfully, the script is probably a syllabary like Linear A and B. In a syllabary, most signs represent syllables, whereas in an alphabet the signs represent vowels or consonants. Syllabaries use more signs than alphabets: 48 for a Japanese *kana* and 87 for Linear B, for instance. The 45 signs on the Phaistos Disc are too numerous for any known alphabet; the largest, Russian, has 36. And they are far too few to resemble a script such as Egyptian hieroglyphic or Babylonian cuneiform, which boast hundreds of logograms (word signs) along with their core phonetic signs. Moreover, the length of the disc's sections supports a syllabary — such scripts typically have words of this length as syllabaries are more concise than alphabets.

The full script probably used more signs than appear on the disc. A small sample of a text might omit less frequent signs: the preceding paragraph, for example, contains no 'q', 'x' or 'z'. Linguists have a formula for calculating the probable number of signs in an alphabet or syllabary from a small text sample. It works well with modern languages and writing systems such as the English alphabet, the Arabic consonantal script and the Japanese syllabic *kana*, and also with Linear B. This formula predicts a syllabary of 56–57 signs when applied to the characters of the Phaistos Disc, as Alan Mackay demonstrated in 1965. So there were probably 11 or 12 more signs than we see on the disc. This total would be manageable for printing — unlike, say, hundreds of logograms.

Every successful decipherment of an ancient script, from Egyptian hieroglyphs in the 1820s through Linear B in the 1950s, up to the Mayan glyphs of the past few decades, has depended for general acceptance on testing against plentiful virgin inscriptions. At present, all leading Minoan script researchers are compelled to concede that to make further progress on the Phaistos Disc we must hunt for more examples around the shores of the eastern Mediterranean. Such a breakthrough occurred with another fascinating solitary inscription — that of the Tuxtla Statuette, found in Mexico in 1902 and sent to the Smithsonian Institution in Washington DC. In 1986 a much more substantial example of the same script turned up at La Mojarra, not far from Tuxtla. The subsequent, controversial, decipherment made the cover of *Science*.

In the meantime, a thermoluminescence test for the Phaistos Disc is imperative. It will either confirm that new finds are worth hunting for, or it will stop scholars from wasting their effort. ■

Andrew Robinson is a visiting fellow of Wolfson College Cambridge, Barton Road, Cambridge CB3 9BB, UK. He is author of *The Story of Writing*, *The Man Who Deciphered Linear B* and *Lost Languages: The Enigma of the World's Undeciphered Scripts*.

NEWS & VIEWS

MEDICAL IMAGING

A colourful future for MRI

Richard Bowtell

Following multiple physiological variables or cell types *in vivo* requires specific probes. Microfabricated magnetic particles could produce such tuneable contrast agents for magnetic resonance imaging.

Optical imaging routinely uses multicoloured contrast agents ranging from traditional chemical dyes and fluorophores to specially engineered quantum dots. In magnetic resonance imaging (MRI), contrast agents have also proved extremely useful, but their effects are largely indistinguishable from one another, leading to essentially monochrome contrast based on increased or decreased signal strength. Now Zabow and colleagues (page 1058 of this issue[1]) are bringing 'colour' to MRI. They have developed an approach to produce MRI contrast agents with characteristic spectral signals, based on the control of mechanical structures. Different versions of these probes can be used simultaneously, and are distinguishable by the geometry-dependent spectral 'colour' of their signals.

The development of chemically synthesized contrast agents, which can highlight specific tissue regions, has helped MRI become the pre-eminent imaging technique in clinical diagnosis and a powerful tool for biomedical research. Contrast agents for MRI act by causing a localized disturbance of the applied magnetic field, which dictates the radio frequency at which nuclear magnetic resonance (NMR) signals occur. Many nuclei or chemical species can be imaged, but MRI usually relies on the NMR signal from the hydrogen nuclei in water molecules.

The most widely used MRI contrast agents[2] are metal complexes with paramagnetic properties. When placed in a magnetic field, these create magnetic fields of their own, which fluctuate in time, resulting in faster relaxation of nearby water molecules to equilibrium after each radio-frequency excitation, and an enhanced NMR signal. Small iron oxide particles, which locally attenuate the signal, are also being used in clinical and research studies — for example, in the tracking of labelled stem cells[3]. With these agents, the static but spatially varying field produced around each magnetic particle reduces the signal, increasing the contrast with the background.

In a fundamentally different approach, Zabow and colleagues[1] use magnetizable particles containing a cavity, inside which a substantial, spatially uniform magnetic field is generated whose strength is significantly different from that of the applied field outside the particle. The magnitude of this difference depends on the particle's shape, but is independent of its absolute size. Water molecules inside the cavity can be excited by appropriately tuned radio-frequency pulses, which have no effect on water outside the cavity or inside particles of different geometry, because the field strength is completely different in these other locations. Unlike conventional particulate agents, which rely on contrast resulting from the relatively uncontrolled variation in field (and so resonant frequency) outside the particle, the engineered particles produce targeted contrast based on direct manipulation of the signal from water inside particles of a specific geometry.

If it were possible to modulate the signal only by targeting water molecules in cavities at the instant that radio-frequency pulses were applied, unfeasibly high particle densities would be needed to produce useful effects. However, Zabow *et al.*[1] use open structures that water molecules can diffuse in and out of very rapidly. In this situation, repeated application of tuned radio-frequency pulses modulates the signal from a volume of water that is many times larger than that contained within the particles. This diffusional enhancement allows a low density of micro-engineered particles to produce useful contrast.

To demonstrate their approach, the authors used surface micro-machining to form nickel particles with an open, double-disk geometry (Fig. 1). Inside an MRI scanner, the particles align with the applied magnetic field, and a uniform, negative field offset — the magnitude of which depends on the disks' geometry, spacing and saturation magnetization — is produced in the space between the disks. Particles with sizes varying from a couple of micrometres to just over a millimetre generate geometry-dependent resonant frequency shifts up to almost 400 kilohertz for hydrogen nuclei. This demonstration of controlled modulation of signal intensity by targeted radio-frequency irradiation included the formation of a coloured 'red–green–blue' image showing the distributions of three different particle types present in a sample (see Fig. 3b on page 1060).

This work illustrates the potential of microfabricated contrast agents for MRI, but there is much to be done before these structures can be

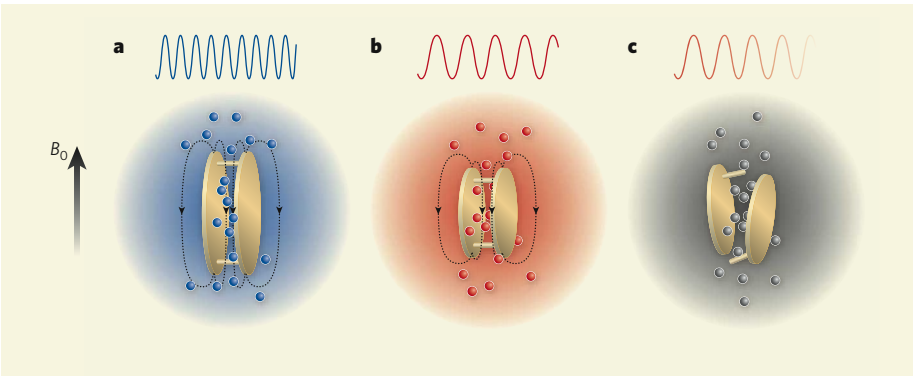


Figure 1 | Particulate contrast agents. a, Particles consisting of gold-coated nickel disks separated by non-magnetic spacers align in a magnetic resonance imaging (MRI) scanner's applied magnetic field (B_0). Magnetization of the nickel generates a magnetic field in opposition to the applied field in the gap between the disks (dashed field lines). **b**, Particles of different geometry (radius, thickness and spacing) create different field offsets and so different changes in the nuclear magnetic resonance (NMR) frequency of the enclosed water. Diffusion of molecules in and out of the gap increases the volume of water affected by repeated application of radio-frequency pulses. **c**, Particles could be functionalized by using spacers that dissolve under specific physiological conditions, eliminating the contrast signal.

routinely used *in vivo*. This includes the development of methods allowing targeted delivery of millions of accurately sized, biocompatible structures and the reduction of particle dimensions to less than a micrometre. But there do not seem to be any insurmountable barriers on the road to exploitation.

There is already an alternative approach for generating controlled spectral shifting based on the use of molecular complexes containing paramagnetic ions[4]. These complexes, known as PARACEST agents, have previously been shown to be capable of monitoring changes in physiological state. But back-of-the-envelope calculations[1] indicate that microfabricated particles will generate usable contrast at much lower agent concentrations. In addition, the large range of continuously variable frequency offsets that can be produced by magnetic structures, even on relatively low-field scanners, will be advantageous compared with the smaller, more discrete shifts produced by macromolecule-based agents.

The controlled spectral signatures provided by microfabricated particles can be immediately used in various applications, including cell tracking and microfluidics. But the ultimate usefulness of these agents *in vivo* will probably depend on the feasibility of making magnetizable particles sensitive to physiological conditions. Zabow *et al.*[1] suggest several ways in which this could be done through reversible or irreversible changes in particle geometry aimed at modulating the field offset in the cavity. Interdisk spacers in double-disk particles could, for example, dissolve (Fig. 1) or change in length in response to particular physiological conditions, such as pH or the presence of specific metabolites. If these functionalities can be realized in living systems, the resulting multispectral, physiologically sensitive contrast would add real colour to MRI's bright future. ■

Richard Bowtell is at the Sir Peter Mansfield Magnetic Resonance Centre, School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK.
e-mail: richard.bowtell@nottingham.ac.uk

1. Zabow, G., Dodd, S., Moreland, J. & Koretsky, A. *Nature* **453,** 1058–1063 (2008).
2. Caravan, P., Ellison, J. J., McMurry, T. J. & Lauffer, R. B. *Chem. Rev.* **99,** 2293–2352 (1999).
3. Bulte, J. W. & Kraitchman, D. L. *NMR Biomed.* **17,** 484–499 (2004).
4. Zhang, S., Winter, P., Wu, K. & Sherry, A. D. *J. Am. Chem. Soc.* **123,** 1517–1518 (2001).

NEUROSCIENCE

Brain control of a helping hand

John F. Kalaska

Paralysed patients would benefit if their thoughts could become everyday actions. The demonstration that monkeys can use brain activity for precise control of an arm-like robot is a step towards that end.

Strokes, spinal-cord injuries and degenerative neuromuscular disease all cause damage that can severely compromise the ability of patients to use their muscles. The loss of mobility and independence that results from such motor deficits takes a devastating toll on their quality of life. Medical research is striving on many fronts to reverse the disease or injury state of such patients. Meanwhile, other approaches are needed to enhance their quality of life. Research described by Velliste *et al.*[1], published on page 1098 of this issue, provides a heartening example of what, in due course, may be possible*.

Often, the patient's condition leaves intact parts of the cerebral cortex involved in voluntary motor control, including the primary motor cortex, premotor cortex and posterior parietal cortex. These patients are still able to produce the brain activity that would normally result in voluntary movements, but their condition prevents those signals from either getting to the muscles or activating them adequately. In such cases, one possible solution is to let the subjects think about what they would like to do as if they were mentally rehearsing the desired actions, record the resulting brain activity, and use those signals to control a robotic device. The development of such brain–machine interfaces (BMIs), or neuroprosthetic controllers, is being pursued in several laboratories.

Velliste *et al.*[1] report one of the latest advances in this field. Using grids of fine electrodes implanted in the primary motor cortex of monkeys, they trained the animals to generate patterns of brain activity to control an anthropomorphic robot arm that had a shoulder joint, an elbow joint and a claw-like gripper 'hand'. The animal sat with its arms gently restrained at its side, with the robot arm positioned next to its shoulder (see Fig. 1a of the paper[1]). Remarkably, within a few days, the monkeys were able to make the robot reach out to a tasty treat such as a piece of fruit, stop, close the gripper on the treat, remove it from a small peg, bring the treat-laden gripper back to their mouth and open the gripper to eat the treat, all in one natural-looking motion.

This is the first reported demonstration of the use of BMI technology by subjects to perform a practical behavioural act — feeding themselves — via brain control of the motion of a robotic arm in three-dimensional space. It represents the current state of the art in the development of neuroprosthetic controllers for complex arm-like robots that could one day, in principle, help patients perform many everyday tasks such as eating, drinking from a glass or using a tool.

One encouraging finding was how readily the monkeys learned to control the robot. Velliste *et al.*[1] used standard operant conditioning methods, in which each successful reach, grasp and retrieval was reinforced by consumption of the food reward. The initial training period was assisted by corrective signals generated by the BMI control program, but the monkeys quickly learned how to generate the brain activity that would produce the desired robot motions without any assistance. Learning could be even quicker in human subjects, facilitated by verbal instructions from a trainer. This also suggests that neuroprosthetic devices could minimize the frustration often felt by patients in current rehabilitation programmes when their diminished motor capacity results in only small performance gains despite prolonged, intense effort.

Equally encouraging was how naturally the monkeys controlled and interacted with the robot. They made curved trajectories of the gripper through space to avoid obstacles, made rapid corrections in the trajectory when the experimenter unexpectedly changed the location of the food morsel, and even used the gripper as a prop to push a loose treat from their lips into their mouth. The monkeys evidently adapted to the robot as a natural surrogate for their own immobile arm. Previous findings indicated that when monkeys learn to use a tool, it becomes incorporated into their own internal body image[2]. Patients who use neuroprosthetic devices for any period of time may come to regard them as natural extensions of their own body, because they can control them efficiently and relatively effortlessly through their own thought processes. This bodes well for the long-term psychological well-being of patients who must depend on such technology.

Velliste *et al.*[1] have provided a promising further proof-of-concept of the potential of BMI technology to help neurological patients. However, we should not get carried away and leap to the conclusion that neuroprosthetic robots will soon be available at the local rehabilitation clinic. All of the main technology employed by Velliste *et al.* to use brain activity to control an arm-like robot has already been demonstrated with simpler remote devices, first in experimental animals[3–7], and recently in human clinical subjects[8]. They offer no fundamental conceptual or technical advances to surmount several hurdles that must still be overcome to permit the widespread clinical application of neuroprosthetic control technology.

For example, the long-term reliability of the implantable electrodes must be improved.

*This News & Views article and the paper concerned[1] were published online on 28 May 2008.

Patients will need to use this technology for many years, but the quality of the recorded neural activity often deteriorates within weeks or months. Furthermore, the success of neuroprosthetic control has been confined so far to the laboratory environment, because the current technology involves a sizeable array of relatively immobile recording, computer and robotic control hardware whose operation also requires the constant attention of a skilled technician. Much work remains to be done if neuroprosthetic controllers are to become portable and largely autonomous.

Moreover, to date, subjects have used only visual feedback to control remote devices. For physical interactions with the environment, the subjects must also be able to sense and control the forces exerted by the robot on any object or surface — so that, for instance, they can pick up an object with a strong enough grip to prevent it slipping from the robotic hand but not so strong as to crush it. This vital information is normally provided by sensory receptors in the skin, muscles and joints. The robots must be equipped with equivalent sensors, and some efficient means must be developed to deliver this sensory feedback[9] to the patients. These and other technical issues are challenging, but not insurmountable.

Like many previous BMI studies[3–7], Velliste *et al.*[1] recorded from the primary motor cortex. Other studies have extracted potential control signals from the premotor cortex and the posterior parietal cortex[7,10–12]. Signals from each of these brain regions have unique properties that may make them particularly useful for different aspects of voluntary behaviour. This may lead to the eventual development of 'intelligent' neuroprosthetic controllers. Such controllers would allow patients with severe motor deficits to interact and communicate with the world not only by the moment-to-moment control of the motion of robotic devices, but also in a more natural and intuitive manner that reflects their overall goals, needs and preferences[10,11]. ■

John F. Kalaska is in the Département de Physiologie, Groupe de Recherche sur le Système Nerveux Central, Université de Montréal, Montréal, Québec H3C 3J7, Canada.
e-mail: kalaskaj@physio.umontreal.ca

1. Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S. & Schwartz, A. B. *Nature* **453,** 1098–1101 (2008).
2. Maravita, A. & Iriki, A. *Trends Cogn. Sci.* **8,** 79–86 (2004).
3. Chapin, J. K., Moxon, K. A., Markowitz, R. S. & Nicolelis, M. A. L. *Nature Neurosci.* **2,** 664–670 (1999).
4. Wessberg, J. *et al. Nature* **408,** 361–365 (2000).
5. Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R. & Donoghue, J. P. *Nature* **416,** 141–142 (2002).
6. Taylor, D. M., Tillery, S. I. & Schwartz, A. B. *Science* **296,** 1829–1832 (2002).
7. Carmena, J. M. *et al. PLoS Biol.* **1,** e42 (2003).
8. Hochberg, L. R. *et al. Nature* **442,** 164–171 (2006).
9. London, B. M., Jordan, L. R., Jackson, C. R. & Miller, L. E. *IEEE Trans. Neural Syst. Rehabil. Eng.* **16,** 32–36 (2008).
10. Musallam, S., Corneil, B. D., Greger, B., Scherberger, H. & Andersen, R. A. *Science* **305,** 258–262 (2004).
11. Hatsopoulos, N., Joshi, J. & O'Leary, J. G. *J. Neurophysiol.* **92,** 1165–1174 (2004).
12. Santucci, D. M., Kralik, J. D., Lebedev, M. A. & Nicolelis, M. A. *Eur. J. Neurosci.* **22,** 1529–1540 (2005).

CANCER

Deconstructing oncogenesis

Ji Luo and Stephen J. Elledge

Transformation of normal cells into cancer cells entails concerted changes in the expression of many genes. Identifying which of those genes are crucial will provide insight into the mechanisms underlying malignancy.

Multiple genetic alterations pave the way for transformation of a normal cell into a cancer cell. At the core of this process are oncogenic (cancer-causing) mutations in critical genes, which lead to sustained proliferative drive and desensitization of the cell to cues that normally inhibit growth or promote cell death[1]. Almost exactly 25 years ago, two landmark papers[2,3] published in *Nature* showed that a single oncogene is insufficient to make normal cells cancerous, whereas cooperation between two distinct oncogenic mutations can do the job. Writing on page 1112 of this issue[4], some of the same authors now delve deeper into the genetics of oncogene cooperation and uncover many genes whose altered expression is important for malignancy. Intriguingly, in many cases 'normalizing' the expression of even one of these genes is sufficient to attenuate tumour growth.

McMurray *et al.*[4] studied the cooperation between two classic, highly prevalent oncogenic mutations in human cancers — that of Ras and of p53. The small GTPase proteins of the Ras family are signalling molecules normally activated by growth-factor receptors to promote cell proliferation. Single-nucleotide mutations that lead to the constitutive activation of Ras are found in more than 30% of human cancers[5]. The tumour suppressor protein p53 is a gene transcription factor that is normally activated by stress such as DNA damage, oxygen shortage or the presence of oncogenes (such as mutant Ras) to halt cell proliferation and promote programmed cell death (apoptosis). In most human cancers, p53 function is compromised owing to mutations in either p53 itself or elsewhere in the p53 signalling pathway[6]. Whereas oncogenic mutations in either Ras or p53 have some effect on turning normal cells malignant, together these mutations create the perfect storm of malignant transformation that readily promotes tumour growth[2].

One hypothesis to explain oncogene cooperation is that different oncogenic mutations might join forces to increase the expression of genes that promote tumour formation and downregulate those that suppress this process. To test this hypothesis, McMurray *et al.* compared gene-expression profiles in mouse colon cells that expressed no oncogene; only a constitutively active Ras oncogene; only a mutant p53; or both mutant Ras and p53. Among the hundreds of genes whose expression changed when these mutants were introduced individually or together, the authors identified a subset of 95 that they refer to as "cooperation response genes" (CRGs); the expression of CRGs was synergistically increased (28 genes) or decreased (67 genes) by mutant Ras and p53 (Fig. 1).

But do CRGs have a critical role in supporting the malignant phenotype of cells expressing mutant Ras and p53? The authors randomly selected 24 CRGs and, one at a time, painstakingly restored their expression levels

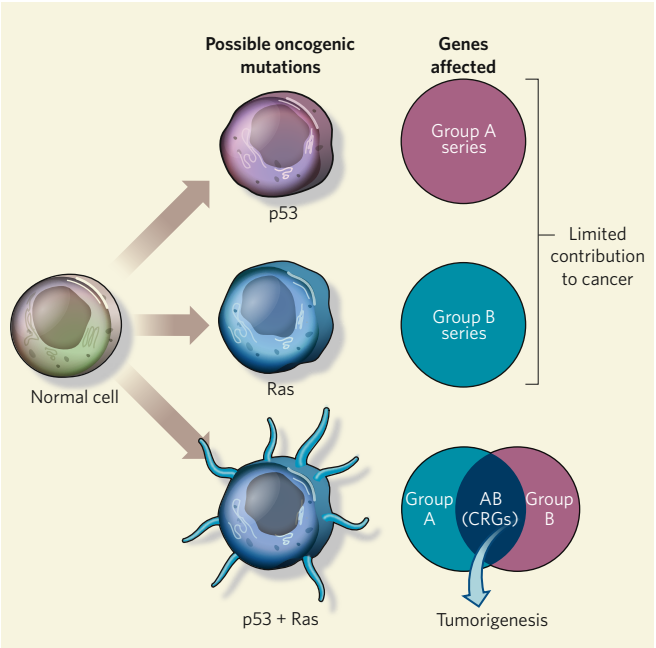


Figure 1 | Cooperation response genes. Oncogenic mutations in the transcription factor p53 and in the small GTPase protein Ras — which individually have limited effects on promoting cancer — cooperate to transform normal cells into cancer cells[2]. In this example, p53 mutation affects the expression of group A genes and Ras mutation modifies the expression of group B genes. When both p53 and Ras are mutated in the same cell[4], they synergistically regulate a subset of genes (AB) known as cooperation response genes (CRGs), which turn out to be crucial mediators of tumour formation.

to that seen in normal cells through either gene overexpression or RNA-interference (RNAi)-mediated depletion of gene expression. As a control, they also individually normalized the expression of 14 non-CRGs — that is, genes whose expression was perturbed by either mutant Ras or p53, but not synergistically by both. Surprisingly, for a high proportion (58%) of the 24 CRGs, restoring normal expression levels attenuated tumour growth when the cells were transplanted into mice. By contrast, only one of the 14 non-CRGs behaved similarly. So CRGs, many of which are implicated in cancer for the first time in this study[4], seem to be highly enriched in genes involved in promoting tumour formation.

McMurray and colleagues' CRGs, identified using the mouse colon cells, are also likely to be important in human cancers that involve mutations in Ras and p53. The authors find that more than half of the CRGs tested show similarly altered expression in specimens of human colon cancer. Furthermore, manipulating the expression of several CRGs in human colon-cancer cells with mutations in p53 and Ras pathways also attenuated tumour growth when these cells were transplanted into mice.

Although McMurray and colleagues' findings indicate that the effect of oncogene cooperation depends in part on CRGs, it is possible that non-CRGs also have an as yet unclear role in this process. Moreover, it will be important to determine whether other oncogene pairs similarly regulate overlapping sets of genes that are functionally crucial for tumour formation.

The CRGs McMurray *et al.* describe encompass genes with diverse cellular functions, including signal transduction, transcription, apoptosis and cellular metabolism. This provides a plausible explanation for how a few oncogenic mutations in 'master regulators' such as Ras and p53 could initiate broad cancer-associated changes. They do so by mobilizing genetic programmes that coordinate many aspects of cellular function essential for malignant transformation. For example, the metabolic programme might serve to increase biosynthesis for rapid cell division; changes in the apoptotic programme might promote cell survival; and the signalling and transcription programmes might free the cell from the constraints of external regulatory cues.

A central issue in understanding tumour formation is determining the number of genes that actually contribute to this process. If this study's results[4] can be generalized, it is likely that hundreds of genes each make partial contributions to the cancer phenotype. Indeed, the complex alterations observed in cancer genomes[7], and the finding[8] that mutations in almost 20% of genes encoding protein kinase enzymes might contribute to the oncogenic state, both support this possibility. Given the extreme rewiring of genetic networks in tumours, at present there is no logical way to predict which genes — oncogenes and non-oncogenes alike[9,10] — will make a sufficiently important contribution to tumour maintenance to represent viable drug targets. Whereas oncogenes can be identified through genome analyses, the discovery of tumour 'addiction' to non-oncogenes requires alternative approaches. It is likely that the use of gain-of-function (overexpression) or loss-of-function (RNAi) approaches to interrogate the dependencies of cancer cells, either through testing candidate genes, as demonstrated by McMurray *et al.*, or by genome-wide screens, will be the least biased path to deconstructing the vulnerabilities of cancer cells for developing therapies. ■

Ji Luo and Stephen J. Elledge are in the Howard Hughes Medical Institute and the Department of Genetics, Center for Genetics and Genomics, Harvard Medical School, Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.
e-mail: selledge@genetics.med.harvard.edu

1. Hanahan, D. & Weinberg, R. A. *Cell* **100,** 57–70 (2000).
2. Land, H., Parada, L. F. & Weinberg, R. A. *Nature* **304,** 596–602 (1983).
3. Ruley, H. E. *Nature* **304,** 602–606 (1983).
4. McMurray, H. R. *et al. Nature* **453,** 1112–1116 (2008).
5. Downward, J. *Nature Rev. Cancer* **3,** 11–22 (2003).
6. Toledo, F. & Wahl, G. M. *Nature Rev. Cancer* **6,** 909–923 (2006).
7. Wood, L. D. *et al. Science* **318,** 1108–1113 (2007).
8. Greenman, C. *et al. Nature* **446,** 153–158 (2007).
9. Weinstein, I. B. & Joe, A. K. *Nature Clin. Pract. Oncol.* **3,** 448–457 (2006).
10. Solimini, N. L., Luo, J. & Elledge, S. J. *Cell* **130,** 986–988 (2007).

ORGANIC ELECTRONICS

On the border

Liesbeth Venema

At the interface between two compounds, physical properties can emerge that neither material displays on its own. A striking example of such an effect occurs at the border between two organic molecular crystals.

Sometimes two people seem perfect for each other, but arrangements to get them together just don't work out. Materials scientists face a similar challenge when wanting to bring together two different compounds in the hope that the combined structure will have useful properties. As they describe in *Nature Materials*, Alves *et al.*[1] have undertaken such a piece of matchmaking — with a happy outcome. They simply pressed two organic molecular crystals together and made the remarkable discovery that, at the interface, an electronic system unlike any other is created.

Separately, the two solids are poor conductors of electrical charge and are classed as semiconductors with a large energy bandgap. But when pieces of the two crystals are stuck together, a conductive region appears between them. Alves and colleagues' detailed electronic measurements indicate the intriguing possibility that a thin metallic system is formed that consists of two separate layers with equal but opposite amounts of charge.

Organic, carbon-based molecules are not generally known as good conductors, but there is nonetheless great interest in their electronic properties. That's because they have several virtues — design flexibility, compatibility with a wide range of substrates and cost-effectiveness — that make them appealing for use in low-cost 'flexible electronics'. One way to make organic material more conductive is to insert impurity atoms that donate mobile charges to their molecular host. Another approach is to mix two carefully chosen molecules that exchange charge with each other and that together form a crystalline, conducting solid — a so-called charge-transfer salt.

Even though most organic electronic devices are made from thin films, little attention has been paid to electronic phenomena at the interfaces between these films. But studies of events at such borders can be exceedingly fruitful, as discoveries in a neighbouring research area show. This is the area of 'oxide electronics', where electronic and magnetic materials are engineered by stacking layers of inorganic transition-metal oxides. Careful control of the atomically smooth interfaces can produce effects that are not observed in any of the individual oxide crystals. A striking example[2] was the formation of a two-dimensional, quantum-mechanical system of freely moving electrons with exceptionally high mobility at the interface of two different oxides that are electrical insulators on their own.

But how might the promise of interfacial effects be exploited for organic electronics? Alves *et al.*[1] turned their attention to two organic molecules that constitute the first example of a charge-transfer salt, and so are already known to get along well[3]. The molecules — TTF (tetrathiofulvalene) and TCNQ (7,7,8,8-tetracyanoquinodimethane) — co-crystallize into a solid, lining up separately into orderly chains (Fig. 1). As a salt, the two different molecules don't form chemical bonds. But TTF donates electrons to TCNQ, and TCNQ in turn offers positive charges, or holes, to TTF. (These holes are 'missing electrons' that behave as mobile charges in their own right.)

Charge-transfer salts are highly conducting at room temperature. But an odd feature emerges when they are cooled. A structural instability

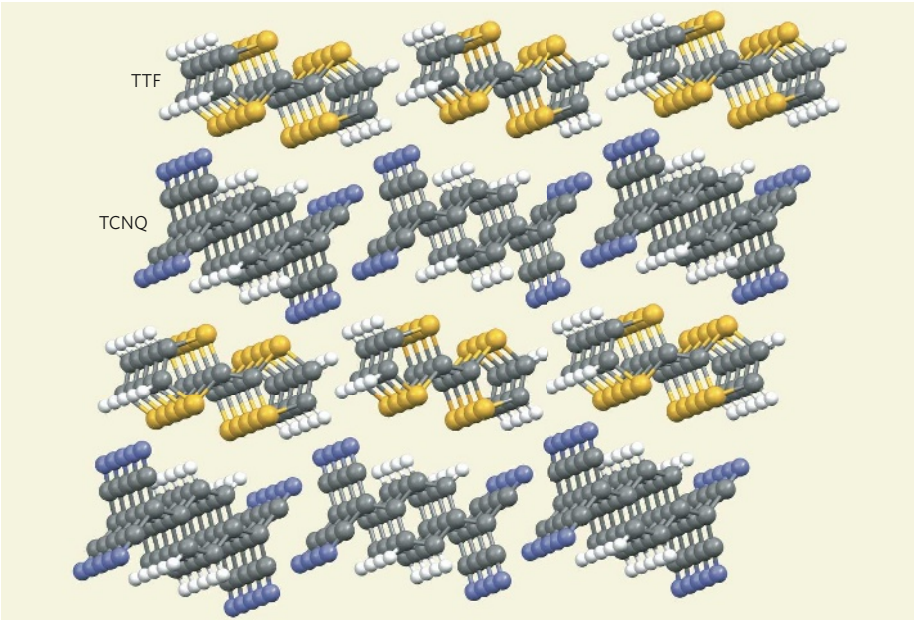


Figure 1 | Organic crystals in chains. In a charge-transfer salt, molecules of TTF (tetrathiofulvalene) and TCNQ (7,7,8,8-tetracyanoquinodimethane) line up into orderly chains to form a crystalline structure. They do not interact chemically but exchange charge, so that the solid is highly conductive at room temperature. Alves and colleagues[1] investigated whether a similar charge transfer occurs not only in the solid but at the interface between TTF and TCNQ crystals. Grey, carbon atoms; yellow, sulphur; blue, nitrogen; white, hydrogen. (Reproduced from ref. 1, courtesy of the Cambridge Structural Database.)

occurs independently in the linear arrangements of both molecules (the Peierls transition, which generally afflicts one-dimensional structures), with molecules within the chains alternately moving a bit closer and a bit farther away from each other. The upshot is that conduction of charge along the chains is impeded and the compound turns into an insulator.

In their experiments, Alves *et al.* were entering unknown territory. It was far from obvious how the molecules would behave when the surfaces of TTF and TCNQ crystals were brought into contact, and whether the charge exchange seen in the TTF–TCNQ charge-transfer salts would occur between the surface layers of the two separate crystals. Other questions were whether the surfaces would fit together smoothly enough (any imperfections would trap mobile charges), and whether diffusion of molecules across the interface would mix the two substances, effectively producing a mini charge-transfer salt.

Alves *et al.*[1] found that the TTF and TCNQ crystals are sufficiently smooth and flexible to be combined using a notably simple assembly process: manually, in air, TCNQ was first placed on a soft polymeric substrate and TTF was subsequently laminated on top. The resulting interface between them is highly conductive, indicating that charge transfer across the interface does indeed take place, and that charge trapping at the surfaces is not a big issue. Moreover, having studied about ten devices in detail, the authors found from the temperature dependence that the interfacial layers behave as a true metal. In particular, there is no hint of a Peierls transition and corresponding onset of insulating behaviour at low temperature — showing that the molecules at the surfaces only exchange electrons and holes across the interface and do not cross the boundary themselves.

Remarkably, it seems that a metallic layer is produced at the border between the two insulating solids. Moreover, the tantalizing possibility is that here we have a new kind of electronic system, consisting of two weakly coupled sheets of electrons and holes separated from each other at an intermolecular distance. Unexpected electronic and optical effects may well occur in such a system, for example owing to the formation of an ordered layer of excitons (coupled electron–hole pairs that can emit light), as correlated behaviour between the charge carriers can come into play.

Further study of the nature of the interfacial metallic system is required, but a potentially wide avenue for work on organic electronics has opened up. The obvious attraction is the refreshingly straightforward fabrication method; similar lamination processes have already been used to make a range of organic single-crystal transistors. So far, it has been the electronic properties of the whole device rather than those of the interfaces that have been the focus of attention. But that looks set to change. ■

Liesbeth Venema is a Senior Editor of *Nature*.
e-mail: l.venema@nature.com

1. Alves, H., Molinari, A. S., Xie, H. & Morpurgo, A. F. *Nature Mater.* doi: 10.1038/nmat 2205 (2008).
2. Ohtomo, A. & Hwang, H. Y. *Nature* **427**, 423–426 (2004).
3. Ferraris, J., Cowan, D. O., Walatka, V. & Perlstein, J. H. *J. Am. Chem. Soc.* **95**, 948–949 (1973).

50 YEARS AGO

The National Science Foundation's report on Government–University Relationships in Federally Sponsored Scientific Research and Development ... directs attention to three major trends [between 1940 and 1958]... In this period Federal support has extended from the agricultural sciences to every field of natural science. Secondly, the period has seen the innovation and expansion of federally owned and financed research centres; and thirdly, in contrast to the relative absence of Federal 'extramural' financial support of research facilities, a significant part of Federal support (265 million dollars in 1957–58) goes for construction or operation of major research facilities.

From *Nature* 21 June 1958.

100 YEARS AGO

La Loi des petits Nombres. By M. Charles Henry. The question discussed by the author may be given in his own words:—

"Est-il possible de prévoir une loi de séquence plus ou moins fragmentaire dans les phénomènes fortuits comme les arrivés de la rouge et de la noire à la roulette?"

He considers that the theory of probabilities is only verified in practice when the number of throws of the ball is indefinitely great, and that new principles are required when the period of play is short. He takes what he terms a psychophysical point of view, and bases his researches on the ultimate vibrations of particles and the musical interval, the fifth —the ratio 3 : 2. He adopts the latter as governing the sequences at roulette without giving any scientific reason whatever.

It is difficult to take the author seriously, but as he pretends in chapter iv. of the work to give rules of play which will enable a player to win at Monte Carlo, it is necessary to inform the reader that the system of M. Henry is not based upon scientific truth, and can have no effect upon his winning or losing.

From *Nature* 18 June 1908.

50 & 100 YEARS AGO

STRUCTURAL BIOLOGY

Modelling collagen diseases

Barbara Brodsky and Jean Baum

Mutations in collagen lead to hereditary disorders such as brittle-bone disease. Peptide models for aberrant collagens are beginning to clarify how these amino-acid replacements lead to clinical problems.

Collagen is the predominant protein in the body defining the mechanical properties of tissues. In many hereditary connective-tissue disorders, collagen's regular repeating sequence of amino acids is disrupted. Short peptide chains have proved to be valuable models in understanding both these pathologies and normal collagens[1]. In the *Journal of the American Chemical Society*, Gauba and Hartgerink[2] report an intriguing peptide model for osteogenesis imperfecta, a dominant hereditary disorder commonly known as brittle-bone disease. The model allowed them to make disease-related mutations in any or all of the three peptide chains that make up collagen.

Collagen's molecular structure consists of three helical polypeptide chains coiled around each other to form a triple helix. The close packing of these chains creates a precise stagger in their alignment and requires that the smallest amino acid, glycine, occupies every third position in each peptide. The sequence must also have a high content of proline and its modified variant hydroxyproline. Some collagens comprise three identical chains, whereas others contain chains of differing amino-acid composition.

Type I collagen, which is composed of two identical chains and a third with a different amino-acid sequence, is the major protein in bone. Osteogenesis imperfecta is caused by mutations in any of the chains (Fig. 1), which change one of the glycines to a larger residue, such as serine. This delays folding of the protein and decreases the stability of the collagen molecules[3]. The folding defects can lead to retention of collagen within cells and their eventual death, and the structural alterations may result in defective binding of other components required for bone formation[4]. The size and complexity of entire collagen molecules make it difficult to interpret the structural basis of changes in stability and folding. Fortunately, small triple-helix peptide models offer the opportunity to vary amino-acid sequences, allowing the molecular details of changes in a mutation site and its interactions to be defined.

The Hartgerink group has developed a system[5] involving the assembly of three different peptides to form a mixed trimer that represents the latest advance in collagen model design. Peptide models have long been an integral component of collagen structural studies[6], beginning in the 1950s with simple polymers of glycine or proline, and progressing through peptides with strictly repeating amino-acid sequences to triple-helical peptides containing sequences from collagen itself. Peptides with glycine as every third residue and a high content of proline and hydroxyproline self-associate to form stable trimers with three identical chains, known as homotrimers. But attempts to use chains with differing sequences did not produce stable mixed molecules, or heterotrimers. Strategies for forming heterotrimeric triple-helical peptides similar to type I collagen have thus far focused on forcing the desired chain composition and alignment through covalent linkages between the peptides[7–9]. Gauba and Hartgerink[2] have now revisited the idea of creating peptides with different sequences that, when mixed, form stable collagen-like trimers. Instead of being held together by covalent bonds, the authors' design relies on electrostatic interactions between the chains to assemble and align the peptides into a triple helix.

In earlier studies, Gauba and Hartgerink[5] developed an optimal design for assembling and stabilizing heterotrimeric collagen-like peptides by using equal amounts of three different peptides: a neutral peptide consisting of ten repeats of a tripeptide made of proline, hydroxyproline and glycine; a positively charged version in which all of the hydroxyprolines are replaced by the basic amino acid lysine; and a negatively charged variant in which the prolines are replaced by the acidic amino acid aspartic acid. The complementary electrostatic charges on the three peptides lead to the formation of stable trimers with one chain of each type aligned as in native collagen.

In the new study[2], the authors have extended their strategy by incorporating segments of type I collagen sequence into the middle of the model peptides. These inserted sequences carry the osteogenesis imperfecta glycine-to-serine mutation in none, one, two or all of the triple helix's chains (Fig. 1). The positive, negative and neutral sequences flanking the regions of collagen sequence are still able to drive assembly of the peptides into collagen-like heterotrimers.

The authors' measurements of trimer-to-monomer thermal transitions show that a single mutation substantially reduces the stability of these peptides, whereas subsequent mutations lead to less drastic changes. Their experimental results are in good agreement with computational studies[10], which calculated the effect of introducing such mutations into a collagen triple helix. When the authors studied the folding of the peptide trimers, they saw the same trend; the first glycine mutation creates the biggest delay in triple-helix folding. Using Gauba and Hartgerink's approach, it should also be possible to vary the sequence around the mutation site within these peptides, as well as the identity of the amino acid replacing glycine, to investigate factors affecting the severity of osteogenesis imperfecta.

Many other disorders also arise from glycine mutations in collagen, although the clinical symptoms depend on the function and location of the type of collagen harbouring the mutation. For example, Alport syndrome results in progressive kidney disease due to mutations in type IV collagen, whereas Ehlers–Danlos syndrome type IV affects blood vessels as a result of mutations

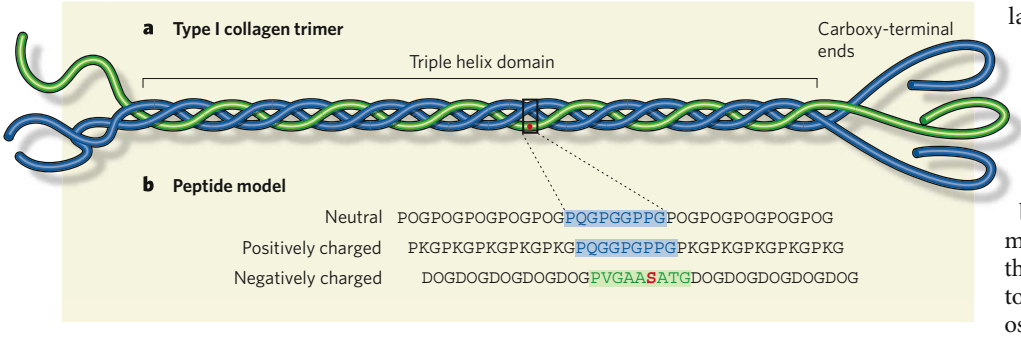


Figure 1 | Collagen mutations mimicked by peptide models. a, Type I collagen is a trimer, composed of two protein chains with identical amino-acid sequences (blue) and one chain with a different sequence (green). Non-triple-helical sequences at the carboxy-terminal ends direct the chain assembly of the molecule. Mutations of the amino acid glycine to serine (red dot) within the helical region result in osteogenesis imperfecta (OI), a dominant genetic bone disease. **b**, Peptides designed by Gauba and Hartgerink[2] incorporate a region of type I collagen sequence between flanking regions with repeating sequences that are positively charged, negatively charged or neutral. The repeating seqences direct trimer formation through electrostatic interactions between the chains. A glycine-to-serine mutation (red) mimics OI. Peptide sequences are given in single-letter amino-acid code, with O used to represent hydroxyproline.

in type III collagen. All such diseases could, in principle, be investigated using the authors' approach.

The techniques used by Gauba and Hartgerink, based on circular dichroism spectroscopy, do not provide high-resolution structural information. But other self-assembling homotrimer triple-helical peptides have proved amenable to molecular-level studies by nuclear magnetic resonance and X-ray crystallography[1,11]. It will therefore be exciting to see whether these self-assembling heterotrimeric peptides[2] can be used to directly visualize the structural perturbation in a collagen disease and provide a basis for rational design of therapeutic drugs. These more realistic peptide models of collagen could also reveal how mutations affect the formation of higher-order structures and interactions with the other components of bone. ■

Barbara Brodsky is in the Department of Biochemistry, University of Medicine and Dentistry of New Jersey — Robert Wood Johnson Medical School, Piscataway, New Jersey 08854, USA. Jean Baum is in the Department of Chemistry and Chemical Biology, BioMaPs Institute, Rutgers University, Piscataway, New Jersey 08854, USA.
e-mails: brodsky@umdnj.edu; jean.baum@rutgers.edu

1. Baum, J. & Brodsky, B. *Curr. Opin. Struct. Biol.* **9,** 122–128 (1999).
2. Gauba, V. & Hartgerink, J. D. *J. Am. Chem. Soc.* **130,** 7509–7515 (2008).
3. Makareeva, E. *et al. J. Biol. Chem.* **283,** 4787–4798 (2008).
4. Marini, J. C. *et al. Hum. Mutat.* **28,** 209–221 (2007).
5. Gauba, V. & Hartgerink, J. D. *J. Am. Chem. Soc.* **129,** 15034–15041 (2007).
6. Brodsky, B. & Persikov, A. V. *Adv. Prot. Chem.* **70,** 301–339 (2005).
7. Slatter, D. A., Foley, L. A., Peachey, A. R., Nietlispach, D. & Farndale, R. W. *J. Mol. Biol.* **359,** 289–298 (2006).
8. Fiori, S., Saccà, B. & Moroder, L. *J. Mol. Biol.* **319,** 1235–1242 (2002).
9. Koide, T., Nishikawa, Y. & Takahara, Y. *Bioorg. Med. Chem. Lett.* **14,** 125–128 (2004).
10. Mooney, S. D., Huang, C. C., Kollman, P. A. & Klein, T. E. *Biopolymers* **58,** 347–353 (2001).
11. Bella, J., Eaton, M., Brodsky, B. & Berman, H. M. *Science* **266,** 75–81 (1994).

EVOLUTIONARY BIOLOGY

The amphioxus unleashed

Henry Gee

The genome sequence of a species of amphioxus, an iconic organism in the history of evolutionary biology, opens up a fresh vista on the comparative investigation of chordates and vertebrates.

One might be forgiven for never having heard of the amphioxus, a small, vaguely fish-shaped creature (Fig. 1), which spends most of its life buried in sand filtering detritus from seawater. Yet for many decades, beginning in the mid-1800s, it was central to a preoccupation with the origin of the vertebrates, the group of backboned animals that includes ourselves. Although lacking a distinct head, organs of special sense or paired fins, the amphioxus has a dorsal tubular nerve cord, and the stiffening axial rod known as the notochord, that are defining features of chordates — the wider group to which vertebrates belong. For much of the twentieth century, the amphioxus was neglected as a subject of study. But with Putnam and colleagues' publication[1] on page 1064 of the draft genome sequence of *Branchiostoma floridae*, one of the 25 or so recognized species of amphioxus, this eldritch organism is set to re-enter public life.

The amphioxus was originally described by P. S. Pallas in 1774 as a kind of slug. It took almost another hundred years before Alexander Kowalevsky recognized the chordate affinities of this organism[2] as well as of tunicates (sea squirts)[3] — the other group of invertebrate chordates — and the golden age of the study of vertebrate evolution began.

In those days, the amphioxus was seen as a vertebrate writ small, bearing clues to our own lost ancestry. No issue of the *Quarterly Journal of Microscopical Science* seemed complete without a study on amphioxus anatomy or development from luminaries such as E. Ray Lankester[4], so that, by 1932, E. G. Conklin[5] felt it necessary to preface a weighty treatise on amphioxus embryology with an apology to his readers for having to wade through yet another substantial work on this creature. But the question of the origin of vertebrates fell from favour because of its abiding intractability, and because of the arrival of genetics and of model organisms such as fruitflies that are easier to study in the laboratory.

The amphioxus was never abandoned, however. In recent years the flame has been kept alive by researchers such as Nicholas and Linda Holland of the Scripps Institution of Oceanography in San Diego, as well as (the unrelated) Peter W. H. Holland at the University of Oxford and an increasing band of students and colleagues. The age of genomics has rescued the amphioxus from chthonic obscurity, as new data — now including Putnam and colleagues' paper[1] and three companion reports in *Genome Research*[6–8] — have reinvigorated the study of the origin of the vertebrates.

The genome of any species, although informative, is hardly more than a matter of record. Two genomes are more interesting, because comparisons can be made between them. But when one has three or more, one can start to frame rather precise hypotheses about the course of genomic evolution, and ask meaningful questions about the origin of morphological novelties. The 520-megabase genome of *B. floridae* would, therefore, be nothing much more than a curiosity without the comparative context offered by the increasing number of completed or draft animal genomes from humans to sea anemones, and in particular those of the tunicates *Ciona intestinalis*[9] and *Oikopleura dioica*[10]. Such studies reveal the amphioxus genome to be, in fact, of preternatural importance. Recent work[11] showing that the amphioxus is the most basal chordate, and not a close relative of vertebrates as had previously been thought, only increases its importance in our understanding of fundamental features of the chordate ancestral condition.

The draft genome underscores the basal position of the amphioxus (Fig. 2, overleaf), revealing strong patterns of conserved

D. L. GEIGER/SNAP/ALAMY

Figure 1 | The amphioxus — back in public life. This species is *Branchiostoma lanceolatum*.

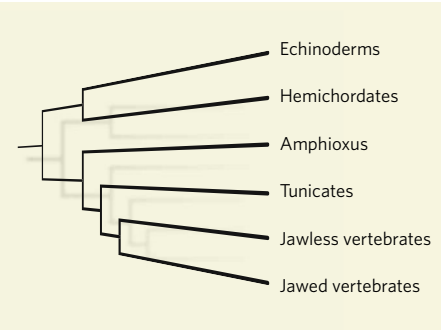


Figure 2 | Chordates and vertebrates. The chordates contain invertebrate groups (amphioxus and tunicates) and vertebrate groups (the jawless vertebrates, such as the lamprey, and the jawed vertebrates or gnathostomes). Amphioxus is the most 'basal' of chordates, a conclusion[11] confirmed by the draft genome sequence of *Branchiostoma floridae*[1]. The closest non-chordate invertebrate relatives of amphioxus are the hemichordates (acorn worms and their allies) and echinoderms (sea urchins and allies).

'synteny' with vertebrate genomes, including the human genome. This shared possession of similar blocks of genes (even though the genes within each block might have been shuffled substantially) is notable, given that the last common ancestor of the amphioxus and vertebrates lived more than 550 million years ago. More remarkable still is the presence of modest amounts of sequence homology between stretches of non-protein-coding DNA in humans and the amphioxus. This information suggests that whatever the common ancestor of all chordates looked like, its genome was similar to that of a modern amphioxus. Such findings also illustrate the degree of morphological and genomic divergence of tunicates from the chordate lineage. Although sequence homology shows tunicates to be more closely related to vertebrates than is the amphioxus[11], their unique pattern of development has been accompanied by dramatic genomic rearrangements and losses of both coding and non-coding stretches of DNA.

This extensive synteny has allowed substantive insight into a suspected episode in the early history of vertebrates when the genome underwent tetraploidization (that is, became quadrupled). Work on the amphioxus shows that this episode — or two closely linked episodes of diploidization, one following hard on the heels of the other — occurred at around the time that the lineage of jawless vertebrates, now represented by forms such as the lamprey (*Petromyzon*), emerged. The extent to which this genomic storm was manifested in the origin of morphological novelty is not known. Yet it is not unreasonable to suggest that it was connected with the origin of gnathostomes —vertebrates with jaws and paired limbs. This is a subject that is little explored as yet, but is likely to be the subject of revelations in coming years, both from genomics and from the discovery of fossil forms. ■

Henry Gee is a Senior Editor of *Nature*.
e-mail: h.gee@nature.com

1. Putnam, N. H. *et al. Nature* **453,** 1064–1071 (2008).
2. Kowalevsky, A. *Mém. Acad. Imp. Sci. Saint-Pétersbourg* **11,** 1–17 (1866).
3. Kowalevsky, A. *Mém. Acad. Imp. Sci. Saint-Pétersbourg* **10,** 1–19 (1866).
4. Lankester, E. R. & Willey, A. *Q. J. Microscop. Sci.* **31,** 445–466 (1890).
5. Conklin, E. G. *J. Morphol.* **54,** 69–151 (1932).
6. Huang, S. *et al. Genome Res.* doi:10.1101/gr.069674.107 (2008).
7. Holland, L. Z. *et al. Genome Res.* doi:10.1101/gr.073676.107 (2008).
8. Yu, J.-K., Meulemans, D., McKeown, S. & Bronner-Fraser, M. *Genome Res.* doi:10.1101/gr.076208.108 (2008).
9. Dehal, P. *et al. Science* **298,** 2157–2167 (2002).
10. Seo, H.-C. *et al. Science* **294,** 2506 (2001).
11. Delsuc, F. *et al. Nature* **439,** 965–968 (2006).

SUPERCONDUCTIVITY

Prospecting for an iron age

Paul M. Grant

Different material options for high-temperature superconductivity — conduction of electricity with little or no resistance at 'practical' temperatures — have arrived. Iron compounds are the latest thing.

High-temperature superconductivity is back in the public eye, and with a bang. But as ever with this topic, we must first journey back to 1986 and 1987, and to Georg Bednorz and Alex Müller[1], and Paul Chu and his colleagues[2]. To start with, there was the headline news[1] of the onset of superconductivity in a previously unexplored class of compounds, the copper oxide perovskites, or layered cuprates, at the then record-setting temperature of 35 kelvins. Shortly afterwards[2], this transition temperature (T_c) was pushed up to 90 K — beyond the temperature of liquid nitrogen.

The initial announcement prompted practically every superconductivity centre on the planet, including my own home lab at IBM Almaden, to ransack the periodic table hoping to strike pay dirt again. So frantic became the search that Tom Lehrer's 1950s classic *The Elements* was chosen as the theme song for a 1988 BBC Horizon documentary, *Superconductor — Race for the Prize*. Special attention was paid to oxides of the first-row transition metals, running from scandium to zinc, with iron compounds among them, in the hope that a companion to copper would emerge. Alas, although the ensuing years saw the transition temperatures climb to 135 K by 1993 (165 K under pressure), that hope remained unfulfilled. The central cation remained copper, complexed in a plane of nearest-neighbour oxygen anions. And the dream of superconductivity at anything close to room temperature, around 300 K, has remained just that.

It now seems we should have looked not only at the transition-metal oxides but also at the transition-metal pnictides — compounds that contain elements from group V (now group 15) of the periodic table, such as nitrogen, phosphorus and arsenic. In mid-May of 2006, a Japanese collaboration[3] reported superconductivity with T_c ~5 K in a compound of stoichiometry originally $La(O_{1-x}F_x)FeP$, consisting of alternating layers of lanthanum-series oxyfluorides and tetrahedrally coordinated ferrous pnictide (Fig. 1). By January 2008 the same group[4] had lifted T_c to 26 K on substituting arsenic for phosphorus, and in April that was raised[5] to 43 K, albeit under an applied pressure of 4 gigapascals. In the meantime, the appearance of papers on the preprint server arXiv posted by a collaboration in China sparked rumours of a T_c of 54 K in $Sm(O_{1-x}F_x)FeAs$. Late last month, that group published a paper[6] describing a T_c of 53.5 K in a pnictide with gadolinium in the lanthanum position.

The increase of T_c in the ferrous pnictides from 2006 and its acceleration since January 2008 is reminiscent of the 'hockey stick' graph seen two decades earlier for increasing T_c in the layered cuprates. Over the past four months, the pnictides have spawned unprecedented numbers of submissions to the condensed-matter part of the arXiv site, sometimes three a day. Most of these contributions are theoretical in nature (theory being a much safer pursuit than experiment, at least physically, especially when a synthesis involving arsenic compounds is concerned), bringing to mind a comment by the late Pierre de Gennes. At a conference that followed the discovery of high T_c, de Gennes admitted that all theoreticians have a "clothes closet" of favourite "models, or suits ... used, unused and over-under-sized", and that when some new superconductor is found they will pull one out, try it on, and "see if it fits".

Where do we go from here? Well, to start with we can see several striking similarities between the ferrous pnictides and the layered copper oxide perovskites. First, both are layered systems. Second, the Ln–O–F layers provide 'charge reservoirs' for doping; they also sterically reduce the overall symmetry with respect to the intervening ferrous pnictide transport planes, quite possibly driving 'Jahn–Teller-like' phonon-driven instabilities. Third, both systems are, to varying degrees, spin-correlated,

quasi-two-dimensional, Mott–Hubbard charge transfer antiferromagnetic insulators in their undoped ground state. These last three properties are believed to be key to high-temperature superconductivity, and are about the only criteria on which you can find (almost) universal agreement among those trying to choose between the bespoke fashions hanging in the high-T_c theoretical closet.

However, observe in Figure 1 that the Fe ions, although nominally Fe^{2+}, analogous to Cu^{2+}, are tetrahedrally coordinated relative to the pnictide anions, as opposed to the square-planar symmetry of the copper oxide compounds. In the first-row transition metals — scandium to zinc — there are ten *d*-electron states (five described by orbital momentum), each of which can hold two electrons with one spin up and another down. We can play with these states to build various cationic configurations. A simple yardstick, called Hund's rule, helps build possible combinations in isolated atoms and ions. It says we have to start filling from the bottom, first occupying each orbital with an up-spin and then starting over again with spin-down, until all available *d*-electrons are consumed. Thus Fe^{2+}, with six electrons at large, will result in a ground state one electron in excess of a half-filled Hund's occupation distribution, and Cu^{2+}, with nine electrons to spend, will yield one electron fewer (a 'hole' or effective positive charge) than a filled *d*-orbital shell. Therefore, in a very crude sense, the new FeAs superconductors can be thought of as the electron analogues of the hole-transporting CuO complexes, and both measurements and theoretical studies bear this out.

The real situation is far more complex than just stated, and simple Hund's rule arguments are confounded by symmetry, position and overlap of neighbouring anions (O^{2-}, $As^{-(3-x)}$), and by Coulomb repulsion that tends to separate spins in otherwise 'Pauli-allowed' cation states from the next nearest cation neighbour. The trade jargon for these effects is 'crystal field splitting', 'hybridization' and 'Hubbard U', respectively. You can be assured each of these is currently undergoing intense exploration. All this notwithstanding, the simple Hund's rule picture that the ferrous pnictides and copper oxides are electron–hole 'duals' may not be simply fortuitous. It may be the reason that, after years of intense searching, nickel and cobalt complexes have not yielded high-temperature superconductors (at least not yet).

In fact, with T_c now at 55 K, are these ferrous pnictides truly 'high-temperature superconductors'? Simply answered, we don't know at present. But it is useful to remember that the expression 'high-temperature superconductivity' did not originate with Bednorz and Müller's paper[1] of 1986. Those who, like myself, are of mature years will recall that this description was coined as a result of studies[7,8] in the 1960s that superconductivity mediated by electron–phonon pairing would top out at around 30 K, and compounds showing anything above this value would be referred to as 'high-temperature materials'. Although the mechanism of high T_c in the copper oxide perovskites remains in question, we do have evidence[9] in MgB_2 that electron–phonon coupling can achieve a transition temperature of 40 K. Is 55 K really that much higher?

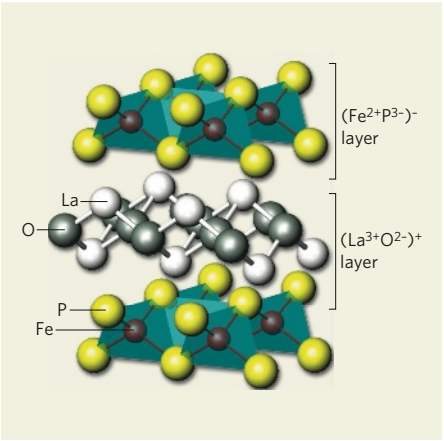


Figure 1 | The unit cell of LaOFeP. In this generic example[3] of the family of lanthanum-series oxyfluoride ferrous pnictides, the overall cell charge is neutral but the individual layers are not, implying electron doping of the FeP layer. Note also that the P coordination of Fe is tetrahedral, not square planar as is the case for the high-T_c copper oxide perovskites. (Reproduced from ref. 3.)

Although most of the theoretical tailoring for the various ferrous pnictides is styled after fashions for the cuprate superconductors, one experimental study harks back to much earlier designs. Chen *et al.*[10] report direct measurement of the superconducting energy gap and its temperature dependence in polycrystalline samples of $Sm(O_{0.85}F_{0.15})FeAs$, with T_c = 42 K. The technique used is called Andreev spectroscopy. This is a variant of tunnelling spectroscopy whereby, at a contact between a normal metal and a superconductor, an electron from the metal injected into the superconductor at energies lower than the superconducting gap gives rise to a superconducting pair (Cooper pair of electrons of opposite spin), which are subsequently spin-charge compensated by a 'reflection' of positive polarity (a hole) back into the normal metal. The resulting current–voltage dependence is a direct measurement of the superconducting pairing energy.

Astonishingly, Chen *et al.*[10] find that their results best fit the time-honoured Bardeen–Cooper–Schrieffer (BCS) theory[11], the breakthrough in the mid-twentieth century that solved the riddle of superconductivity in all materials available up to that time. Although originally formulated to accommodate the pairing of electrons mediated by lattice vibrations (phonons), in its broadest sense the BCS framework can encompass pairing of fermions in a boson field — perhaps even the 'flavours' found in neutron stars, quarks and gluons, giving rise to 'colour' superconductivity at the relatively low cosmological temperature of 10^9 (the units don't matter). So Chen and colleagues' identification of classic BCS behaviour does not rule out the possibility that some more exotic bosonic glue than phonons might be behind superconductivity in these ferrous pnictides.

Whenever a new superconductor with a T_c higher than 30 K appears on the scene, I inevitably get asked if it will bring applications closer. The question is perhaps more pertinent when the material involves particularly noxious elements such as arsenic. My answer is always "Just follow the money." If the pot at the end of the rainbow has enough gold inside (and so far it does not for applied superconductivity), the environmental issues can be overcome. I give you semiconductor manufacture and processing, which uses some of the most toxic compounds (including arsenides) in creation, yet is tolerated and brought under control because its financial return is in the trillions. Again, the units don't matter.

Will T_c in the pnictides continue to go up, and perhaps double or triple as happened in 1987–88? I doubt it. We've now been on standby for several months, and to my mind the best hope is that the discovery of pnictide high-temperature superconductivity will help us understand better the physics of the cuprates. The iron age has yet to dawn. ■

Paul M. Grant is at W2AGZ Technologies, 1147 Mockingbird Hill Lane, San Jose, California 95120, USA.
e-mail: w2agz@pacbell.net

1. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64,** 189–193 (1986).
2. Wu, M. K. *et al. Phys. Rev. Lett.* **58,** 908–910 (1987).
3. Kamihara, Y. *et al. J. Am. Chem. Soc.* **128,** 10012–10013 (2006).
4. Kamihara, Y., Watanabe, T., Hirano, M. & Hosono, H. *J. Am. Chem. Soc.* **130,** 3296–3297 (2008).
5. Takahashi, H. *et al. Nature* **453,** 376–378 (2008).
6. Yang, J. *et al. Supercond. Sci. Technol.* **21,** doi:10.1088/0953-2048/21/8/082001 (2008).
7. McMillan, W. L. *Phys Rev.* **167,** 331–344 (1968).
8. Allen, P. B. & Dynes, R. C. *Phys. Rev. B* **12,** 905–922 (1975).
9. Nagamatsu, J., Nakagawa, N., Muranaka, T., Zenitani, Y. & Akimitsu, J. *Nature* **410,** 63–64 (2001).
10. Chen, T. Y., Tesanovic, Z., Liu, R. H., Chen, X. H. & Chien, C. L. *Nature* doi:10.1038/nature07081 (2008).
11. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. *Phys. Rev.* **108,** 1175–1204 (1957).

Correction

The News & Views article "Genomics: Protein fossils live on as RNA", by Rajkumar Sasidharan and Mark Gerstein (*Nature* **453,** 729–731; 2008), contains the following incorrect statement: "…'reads' found using the Solexa sequencing technology[1,4] can be intersected with some seven pseudogenes, for an average of roughly two reads each." In fact, these reads intersected with some 70 pseudogenes, for an average of roughly 12 reads each. Also, in the text of Box 1, 'nt' (nucleotide) was omitted from one passage, which should read "…to ~27 nt Piwi-interacting RNAs (piRNAs)." These corrections have already been made to the online versions of this article.

OBITUARY

Christopher Curtis (1939–2008)

Medical entomologist and a humanitarian campaigner.

With the death of Chris Curtis on 13 May, the world has lost a leading researcher of insect-borne diseases. A respected theoretical scientist, he was also a pioneer of research into genetic control of disease-carrying insects. But Curtis will probably be best remembered for his innovative contributions to the control of the *Anopheles* mosquito, the vector for the malaria parasite.

He was born in Surrey, UK, and studied at the universities of Oxford and Edinburgh. Early in his research career, he became interested in genetic control of infectious diseases. An example is his seminal work on tsetse flies, the vectors of parasitic African trypanosomes, which, among other diseases, cause sleeping sickness. Curtis used mutation by chromosomal translocation to sterilize these insects, with the idea that they could outcompete the wild, infectious population without the need for radiation or chemical agents. His interest in genetic approaches continued when he moved to India to work for the World Heath Organization (WHO), and he was part of the team that developed sterile male *Culex* and *Aedes* mosquitoes.

Curtis returned to England in 1976 to take a post in the London School of Hygiene and Tropical Medicine, where he found the ambience for research and teaching so ideal that he stayed until his retirement. By the 1980s, he began to realize that genetic approaches were not immediately applicable, and so he focused on practical technologies that would benefit the health of people in developing countries.

In Zanzibar, for example, he demonstrated a low-cost way to control *Culex* mosquitoes, which transmit the parasitic worms responsible for the disfiguring disease filariasis, and the viruses causing West Nile fever and Japanese encephalitis. These mosquitoes survive in pit latrines and soakage pits even after spraying with insecticides. Curtis and colleagues showed that expanded polystyrene beads can form a self-sealing layer on the surface of the water that suffocates the *Culex* larvae, and that a single application of these beads to a pit prevents mosquito breeding for more than seven years. Later, this simple method was used effectively for mosquito control in India, where soakage pits are common sites for mosquito breeding.

Curtis was also an influential figure in the field of malaria management, and the various methods he developed are now considered routine for both intervention and evaluation of *Anopheles* control. In his opinion, the available knowledge in the 1990s of the biology of malaria vectors was sufficient to craft new

A. KOERBER/LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

and affordable malaria control technologies for developing countries, where annual public health budgets were and may still be as meagre as the US$10 per person estimated in 2000.

He was a strong proponent of the use of insecticide-treated nets (ITNs) to cover beds for malaria control, as *Anopheles* mosquitoes bite mainly at night. In collaboration with the National Institute for Medical Research Tanzania, he demonstrated that ITNs are as effective as indoor insecticide sprays for preventing malaria, making these nets central to global malaria control. He also promoted the idea that there can be a 'mass effect' on the mosquito population when everyone in a village — rather than just the vulnerable groups such as pregnant women and children — uses bednets, and that extensive use of ITNs will lead to substantial community protection.

His studies influenced donors and governments. Together with two other campaigners, Awash Teklehaimanot and Jeffrey Sachs, Curtis made a call this year for mass distribution of free, long-lasting ITNs — rather than their allocation through social marketing, whereby each net is sold for US$1–2 — to reduce the burden of malaria. They calculated that an investment of US$3 billion per year, combined with sound public-health measures, would achieve comprehensive malaria control in Africa by 2010. Another important public health lesson was learnt when Curtis and his colleagues suggested that infections can be spread by mosquitoes carried from 'malaria zones' by aeroplanes, thus emphasizing the need for 'disinsection' of air transport.

Curtis continued to investigate genetic approaches to control mosquito populations, but was concerned that the time taken to overcome the hurdles associated with introducing genetically modified insects into the environment would lead to loss of more lives. Indeed, he argued that the current excitement in genomics must be de-emphasized, as far as practical ends are concerned, and that the choice of a molecular method "should be dictated by its being the best way to solve an existing problem" and not by its being the most modern approach. Genetically engineered *Anopheles* strains that cannot transmit malaria would require extremely reliable systems to drive the transgene through the wild population, and that is aside from overcoming objections from society. His other concern was that the better drugs and insecticides designed through genomic approaches would be patent-protected and so too expensive for the poorer nations.

The influence exercised by Curtis, and his achievements, were the result of sound science and meticulous data acquisition. Even in his hobbies he showed that one can have fun while applying scientific rigour. For example, primrose populations were thought to consist of equal numbers of two morphs, but in 1940 the botanist J. Crosby identified a third, self-pollinating, morph — the homostyle — in a Somerset wood and predicted that it would increase over time. Along with his wife Jill, Curtis painstakingly identified and counted primroses of every morph in Somerset to see whether there was a variation in their numbers and distribution over 40 years; they published their findings in the journal *Heredity*.

Chris Curtis will be equally remembered as an inspirational and tireless educator. He was generous as a reviewer of scientific manuscripts, providing extensive suggestions for improving the clarity of manuscripts written by those less skilful in their use of English. An iconic figure for those implementing low-cost vector-control measures, he influenced generations of MSc and PhD students from many countries who now occupy prominent positions. His "So what?" test for new developments and "Don't get it right, get it written!" — his cure for writer's block — became universal lessons for young researchers.

Even though he retired five years ago, he continued teaching and was conducting a five-week vector-control course when he fell ill. He leaves behind his wife Jill and a large community of students and colleagues around the world. My memory is of him in his London laboratory, writing with one hand and providing a blood feed to a beaker full of mosquitoes with the other.

Indira Nath
Indira Nath is in the LEPRA – Blue Peter Research Centre, Cherlapally, Hyderabad 501301, India.
e-mail: indiranath@bprcleprasociety.org

Cover illustration
The entangling of atoms through spin coupling in a double-well potential (Courtesy of I. Bloch)

QUANTUM COHERENCE

Quantum physics has come a long way since its theoretical beginnings in the early twentieth century. Techniques to manipulate light and matter have become increasingly sophisticated, facilitating fundamental studies of quantum effects and inspiring new technologies. From atomic networks to semiconductor 'spintronics', seemingly disparate areas of research are being driven by a shared goal — to harness and exploit quantum coherence and entanglement.

Inevitably, these laboratory endeavours have necessitated a new theoretical toolbox. The image of a pair of photons zooming off in opposite directions, each sensitive to the other through their quantum entanglement, is conceptually tidy. But what happens when describing the quantum properties of more complex systems? This Insight on quantum coherence and entanglement starts with a Progress article that addresses the problem of 'thinking big': how can entanglement be quantified or measured in a system that comprises many particles and degrees of freedom?

The reviews in this Insight highlight the exciting experimental progress in such systems, covering a wide range of physical settings. They describe both bottom-up approaches, in which researchers strive to achieve increasingly complex systems starting from a very small number of particles and degrees of freedom, and top-down approaches, in which the individual and collective degrees of freedom in larger systems are controlled. Ultimately, the goal is to control many-particle systems at the quantum limit, an attractive prospect for quantum simulation and information applications.

As such, this Insight brings together varied research. We trust, however, that you will find coherence in this diversity.

Karen Southwell, Senior Editor

PROGRESS

REVIEWS

nature insight

NATURE|Vol 453|19 June 2008|doi:10.1038/nature07124

Quantifying entanglement in macroscopic systems

Vlatko Vedral[1,2,3]

Traditionally, entanglement was considered to be a quirk of microscopic objects that defied a common-sense explanation. Now, however, entanglement is recognized to be ubiquitous and robust. With the realization that entanglement can occur in macroscopic systems — and with the development of experiments aimed at exploiting this fact — new tools are required to define and quantify entanglement beyond the original microscopic framework.

In the past decade, there has been an explosion of interest in entanglement in macroscopic (many body) physical systems[1]. The transformation in how entanglement is perceived has been remarkable. In less than a century, researchers have moved from distrusting entanglement because of its 'spooky action at a distance' to starting to regard it as an essential property of the macroscopic world.

There are three basic motivations for studying entanglement in the macroscopic world. The first motivation is fundamental. Researchers want to know whether large objects can support entanglement. The conventional wisdom is that a system that consists of a large number of subsystems (for example, 10^{26} of them, similar to the number of atoms in a living room) immersed in an environment at a high temperature (room temperature, for example) ought to behave fully classically. Studying macroscopic entanglement is thus a way of probing the quantum-to-classical transition.

The second motivation is physical and relates to the different phases of matter. Traditionally, the idea of an order parameter is used to quantify phase transitions. For example, a non-magnetic system (in the 'disordered phase') can be magnetized (or become ordered) in certain conditions, and this transition is indicated by an abrupt change in the order parameter of the system. In this case, the magnetization itself is a relevant order parameter, but the interesting question is whether entanglement is a useful order parameter for other phase transitions[2,3].

The third motivation comes from technology. If the power of entanglement is to be harnessed through quantum computing, then entangled systems of increasingly large sizes need to be handled, which is itself a challenge.

It is clear that the modern perspective on entanglement differs greatly from the initial ideas about its seemingly paradoxical nature. Researchers are now realizing how general and robust entanglement is. Larger and larger entangled systems are being manipulated coherently in different physical implementations. And it is not as surprising as it once was to find that entanglement contributes to some phenomena.

Not all of the mystery has vanished, however. As is common in scientific research, answering one question generates many new ones, in this case related to the type of entanglement that is useful for studies motivated by each of the three reasons above. These questions bring researchers closer to the heart of the current understanding of entanglement.

Here I first examine what entanglement is and how it is quantified in physical systems. Different classes of entanglement are then discussed, and I conclude by considering the possibilities of achieving and exploiting large-scale entanglement in the laboratory.

What is entanglement?

The first chapter of almost any elementary quantum-mechanics textbook usually states that quantum behaviour is not relevant for systems with a physical size much larger than their de Broglie wavelength. The de Broglie wavelength, which can intuitively be thought of as the quantum extent of the system, scales inversely as (the square root of) mass

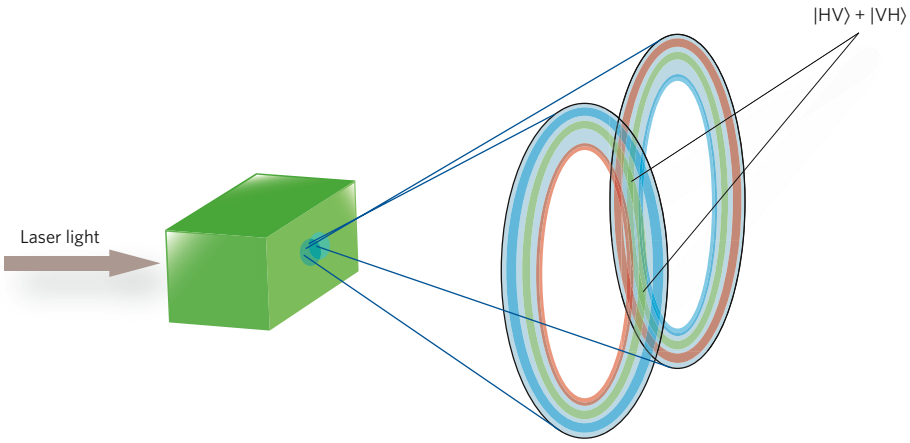


Figure 1 | A way of generating entangled photons by using down conversion. The input laser light is shone onto a nonlinear crystal (green box). The nonlinearity of the crystal means that there is a non-zero probability that two photons will be emitted from the crystal. The cones represent the regions where each of the two photons is emitted. Owing to energy conservation, the frequencies of the photons need to add up to the original frequency. Their momenta also must cancel in the perpendicular direction and add up to the original momentum in the forward direction. One of the photons is horizontally polarized (H), and one is vertically polarized (V). However, in the regions where the two cones overlap, the state of the photons will be $|\mathrm{HV}\rangle + |\mathrm{VH}\rangle$. It is around these points that entangled photons are generated.

[1]School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, UK. [2]Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543. [3]Department of Physics, National University of Singapore, 2 Science Drive 3, Singapore 117542.

times temperature. From this, it can be concluded that massive and hot systems — which could almost be considered as synonymous with macroscopic systems — should not behave quantum mechanically.

As I show in the next section, however, de-Broglie-type arguments are too simplistic. First, entanglement can be found in macroscopic systems[4] (including at high temperatures[5]). And, second, entanglement turns out to be crucial for explaining the behaviour of large systems[6]. For example, the low values of magnetic susceptibilities in some magnetic systems can be explained only by using entangled states of those systems.

Now, what exactly is entanglement? After all is said and done, it takes (at least) two to tangle[7], although these two need not be particles. To study entanglement, two or more subsystems need to be identified, together with the appropriate degrees of freedom that might be entangled. The subsystems are technically known as modes, and the possibly entangled degrees of freedom are called observables. Most formally, entanglement is the degree of correlation between observables pertaining to different modes that exceeds any correlation allowed by the laws of classical physics.

I now describe several examples of entangled systems. Two photons that have been generated by, for example, parametric down conversion[8] are in the overall polarized state $|HV\rangle + |VH\rangle$ (where H is horizontal polarization and V is vertical polarization) and are entangled as far as their polarization is concerned (Fig. 1). A photon is an excitation of the electromagnetic field, and its polarization denotes the direction of the electric field. Each of the two entangled photons represents a subsystem, and the relevant observables are the polarizations in different directions. (Two electrons could also be entangled in terms of their spin value in an analogous way.)

When two subsystems in pure states become entangled, the overall state can no longer be written as a product of the individual states (for example, $|HV\rangle$). A pure state means that the information about how the state was prepared is complete. A state is called mixed if some knowledge is lacking about the details of system preparation. For example, if the apparatus prepares either the ground state $|0\rangle$ or the first excited state $|1\rangle$ in a random manner, with respective probabilities p and $1-p$, then the overall state will need to be described as the mixture $p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1|$. In this case, the probabilities need to be used to describe the state because of the lack of knowledge. Consequently, quantifying entanglement for mixed states is complex.

Systems can also be entangled in terms of their external degrees of freedom (such as in spatial parameters). For example, two particles could have their positions and momenta entangled. This was the original meaning of entanglement, as defined by Albert Einstein, Boris Podolsky and Nathan Rosen[9].

When the subsystems have been identified, states are referred to as entangled when they are not of the disentangled (or separable) form[10]: $\rho_{sep} = \sum_i p_i \rho_1^i \otimes \rho_2^i \otimes \ldots \otimes \rho_n^i$, where $\sum_i p_i = 1$ is a probability distribution and $\rho_1^i, \rho_2^i, \ldots, \rho_n^i$ are the states (generally mixed) of subsystem 1, 2, …, n, respectively. On the one hand, two subsystems described by the density matrix $\rho_{12} = \frac{1}{2}(|00\rangle\langle 00| + |11\rangle\langle 11|)$ are one such example of a separable state. The state of three subsystems, $|000\rangle + |111\rangle$, on the other hand, can easily be confirmed to be not separable and therefore (by definition) entangled.

This simple mathematical definition hides a great deal of physical subtlety. For example, Bose–Einstein condensates are created when all particles in a system go into the same ground state. It seems that the overall state is just the product of the individual particle states and is therefore (by definition) disentangled. However, in this case, entanglement lies in the correlations between particle numbers in different spatial modes. Systems can also seem to be entangled but, on closer inspection, are not (Fig. 2).

Witnesses and measures of entanglement

In this section, I present two surprising results from recent studies of many-body entanglement: first, entanglement can be witnessed by macroscopic observables[11,12] (see the subsection 'Witnessing entanglement'); and second, entanglement can persist in the thermodynamic limit at arbitrarily high temperatures[13]. The first statement is surprising because observables represent averages over all subsystems, so it is expected that entanglement disappears as a result of this averaging. The effect of temperature is similar. Increasing the temperature means that an increasing amount of noise is added to the entanglement, so the second finding — that entanglement can persist at high temperatures — is also surprising.

Before these findings are described in more detail, a simple observation can be made. The entanglement of two subsystems in a pure state is very easy to quantify. This is because the more entangled the state, the more mixed the subset of the system. This property of quantum states — namely that although exact information about the overall state is available, information about parts of the system can be incomplete — was first emphasized by Erwin Schrödinger[14], in the famous paper in which he described the 'Schrödinger's cat' thought experiment.

This logic fails for mixed states, however. For example, an equal mixture of $|00\rangle$ and $|11\rangle$ also results in maximally mixed states for each quantum bit (qubit), but the overall state is not entangled. It also fails for quantifying quantum correlations between more than two components. In fact, in this last case, it is even difficult to determine whether a state of many subsystems is entangled in the first place. This leads on to the concept of witnessing entanglement.

Witnessing entanglement

Entanglement witnesses[15] are observables whose expectation value can indicate something about the entanglement in a given state. Suppose that there is an observable W, which has the property that for all disentangled states, the average value is bounded by some number b,

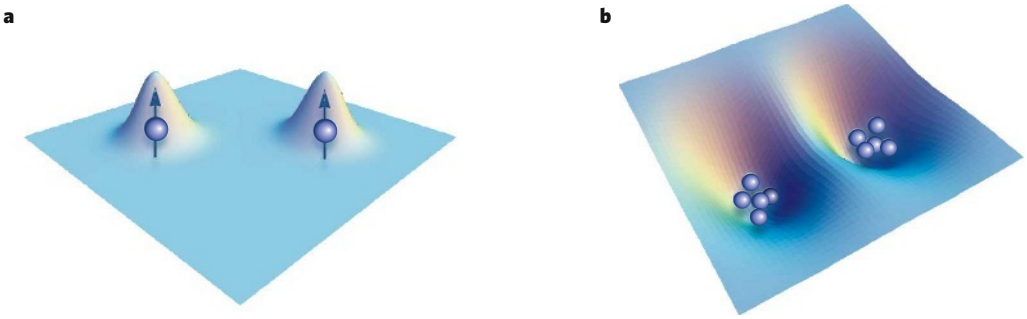


Figure 2 | Separable states. Two examples of disentangled systems are shown. **a**, Two electrons are shown confined to two spatial regions and with their internal spins pointing up. In this case, their spin states are both in the same upwards direction. Because electrons are fermions, the overall state of this system must be antisymmetrical. The internal state is symmetrical (because the electrons are pointing up), and so their special wavefunctions must be antisymmetrized, $|\Psi_1\Psi_2\rangle - |\Psi_2\Psi_1\rangle$. The spatial part of the electronic state, therefore, seems to be entangled — but this only seems to be the case. The electrons in question are fully distinguishable (because they are far apart), so any experiment on one of them is not correlated to any experiment on the other one. Therefore, these electrons cannot be entangled. **b**, A double-well potential, with each well containing five particles, is shown. Experiments that trap atoms in this way are now routine. It is clear that there is no entanglement between the two wells, because each well contains a fixed, clearly defined number of particles, although there could be some entanglement within each well, depending on the exact circumstances.

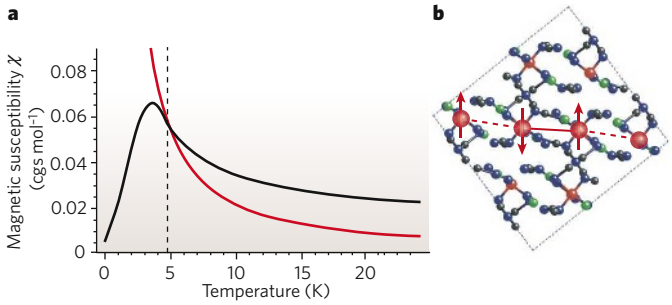


Figure 3 | Susceptibility as a macroscopic witness of entanglement. a, The typical behaviour of magnetic susceptibility χ versus temperature for a magnetic system is depicted (black). The behaviour of the entanglement witness is also depicted (red). Values of magnetic susceptibility below the red line are entangled, and the dashed line indicates the transitional point. **b,** One of the earliest experimental confirmations of entanglement[6] involved copper nitrate, $Cu(NO_3)_2{\cdot}2.5H_2O$, with entanglement existing at less than ~5 K. A molecular image of copper nitrate is depicted, with copper in red, nitrogen in blue, oxygen in black and hydrogen in green. The one-dimensional chain (red), which consists of interacting copper atoms, is the physically relevant chain in terms of the magnetic properties of the compound and can be thought of as a collection of dimers (shown separated by dashed red lines).

$\langle W\rangle \leq b$. Suppose, furthermore, that a researcher is given a physical state and experimentally shows that $\langle W\rangle > b$, then the only explanation is that the state is entangled.

Imagine two spins (a 'dimer') coupled through a Heisenberg interaction[4]: $H = -J\boldsymbol{\sigma}\bullet\boldsymbol{\sigma}$, where H is the hamiltonian, σ denotes a Pauli spin matrix and J is the strength of coupling. I now use the hamiltonian as an entanglement witness. It is easy to see that the average value of H with respect to disentangled (separable) states cannot exceed the value J: $\langle H\rangle = |\mathrm{Tr}(\rho_{\mathrm{sep}}H)| = J|\langle\boldsymbol{\sigma}\rangle\langle\boldsymbol{\sigma}\rangle| \leq J$. However, if the expected values are computed for the singlet state (which is the ground state of H), then the following is obtained: $|\mathrm{Tr}(\rho_{\mathrm{sin}}H)| = 3J$, where ρ_{sin} is the density matrix of the singlet state. This value is clearly outside the range of separable states. The singlet is therefore entangled. This logic generalizes to more complex hamiltonians (with arbitrarily many particles), and it can be shown that observables other than energy (for example, magnetic susceptibility) can be good witnesses of entanglement[1] (Fig. 3). In fact, by using this method, ground states of antiferromagnets, as well as other interacting systems, can generally be shown to be entangled at low temperatures ($k_BT \leq J$, where k_B is the Boltzmann constant and T is temperature; this seems to be a universal temperature bound for the existence of entanglement[16].

Measuring entanglement

Measuring entanglement is complex, and there are many approaches[17]. Here I discuss two measures of entanglement: overall entanglement and connectivity. Further measures are described in ref. 17.

The first measure is overall entanglement, also known as the relative entropy of entanglement[18], which is a measure of the difference between a given quantum state and any classically correlated state. It turns out that the best approximation to the Greenberger–Horne–Zeilinger (GHZ)[19] state, $|000\rangle + |111\rangle$, is a mixture of the form $|000\rangle\langle 000| + |111\rangle\langle 111|$. For W states[20] (by which I mean any symmetrical superposition of zeros and ones, such as $|001\rangle + |010\rangle + |100\rangle$), the best classical approximation is a slightly more elaborate mixture[5].

On the basis of overall entanglement, W states are more entangled than GHZ states. What is the most entangled state of N qubits according to the overall entanglement? The answer is that the maximum possible overall entanglement is $N/2$, and one such state that achieves this (by no means the only one) is a collection of dimers (that is, maximally entangled pairs of qubits). This is easy to understand when considering that each dimer has one unit of entanglement and that there are $N/2$ dimers in total.

The second measure (originally termed disconnectivity) is referred to here as connectivity[21]. Measuring connectivity is designed to address the question of how far correlations stretch. Take a GHZ state of N qubits, $|000\ldots0\rangle + |111\ldots1\rangle$. It is clear that the first two qubits are as correlated as the first and the third and, in fact, as the first and the last. Correlations of GHZ states therefore have a long range. GHZ states have a connectivity equal to N. The W state, by contrast, can be well described by nearest-neighbour correlations. The W state containing $N/2$ zeros and $N/2$ ones can be well approximated by the states $|01\rangle + |10\rangle$ between nearest neighbours. Therefore, correlations do not stretch far, and the connectivity is only equal to 2.

The above considerations of how to quantify entanglement are general and apply to all discrete (spin) systems, as well as to continuous systems (such as harmonic chains[22] and quantum fields[23]), although continuous systems need to be treated with extra care because of their infinite dimensionality. Although the discussed witnesses and measures can be applied to mixed states, I now focus on pure states for simplicity.

Different types of macroscopic entanglement

There are many types of entanglement. Here I discuss the four types that cover all three motivations mentioned earlier: GHZ, W, resonating valence bond (RVB)[24] and cluster[25]. GHZ states are typically used in testing the non-locality of quantum mechanics, because they have a high value of connectivity. W and RVB states naturally occur for a range of physical systems. For example, both Bose–Einstein condensates (such as superfluid and superconducting materials) and ferromagnets have W states as ground states[5].

RVB states are built from singlet states between pairs of spins. It is clear that connectivity of RVBs is only 2, but the states themselves have a high overall degree of entanglement, $N/2$ (ref. 26). It is intriguing that natural states have low connectivity but a high overall entanglement that scales as $\log N$ or even $N/2$, whereas GHZ states, which do not occur naturally, have high connectivity of the order of N but a very low overall entanglement (Box 1).

Are there states that have both connectivity and overall entanglement that scale as the number of subsystems? The answer is, surprisingly, yes. Even more interestingly, these states, which are known as cluster states, are important for quantum computing[25]. Cluster states are highly entangled arrays of qubits, and this entanglement is used to carry out quantum computing through single qubit measurements. Entanglement drives the dynamics of these computers[27], which is why high overall entanglement is needed. But the type of entanglement is also responsible for the implementation of various gates during the operation of these computers, which is why high connectivity is needed.

Experimental considerations and beyond

There are many paths to preparing and experimenting with larger collections of entangled systems. As I have described, natural entanglement is not strong in general and is far from being maximal with respect to overall entanglement or connectivity. To create high overall entanglement and connectivity invariably involves a great deal of effort.

There are two basic approaches to generating large-scale entanglement: bottom up and top down. The first approach, the bottom-up approach, relies on gaining precise control of a single system first and then extending this control to two systems and scaling it up further. So far, 'bottom-up experiments' have obtained up to eight entangled ions in an ion trap (in a W state)[28] and six entangled photons[29]. During nuclear magnetic resonance spectroscopy, 13 nuclei can be 'pseudo-entangled'[30]. Larger systems, however, are exceedingly difficult to control in this way.

The second approach is the top-down approach. As described earlier, many natural systems, with many degrees of freedom (1 million atoms, for example), can become entangled without the need for difficult manipulations (for example, the only requirement might be to decrease the temperature to less than 5 K, which is physically possible). Moreover, in many systems, certain types of entanglement are present in thermal equilibrium and even above room temperature, without the need for any manipulation.

Box 1 | Comparison of four types of entangled qubit state

Qubit state	Overall entanglement	Connectivity
GHZ	1	*N*
W	Log *N*	2
RVB	*N*/2	2
Cluster	*N*/2	*N*

The naturally occurring states, W and RVB, have a much smaller connectivity than the states used for testing non-locality (GHZ) and for carrying out universal computing (cluster). In contrast to connectivity, the overall entanglement shows a different scaling. The important point is that the overall entanglement and connectivity capture markedly different aspects of the 'quantumness' of macroscopic states. Both of these measures can be thought of in terms of fragility of the entangled state, but they describe different types of fragility. The connectivity is related to the fragility of the state under dephasing: that is, the loss of phases between various components in the superposition. The overall entanglement, by contrast, is related to the fragility of the state under the full removal of qubits from the state. For example, if one qubit is removed from the GHZ state, then the remaining qubits automatically become disentangled, which is why the overall entanglement of the GHZ state is equal to 1. If each qubit dephases at the rate *r*, then *N* qubits in GHZ states will dephase at the rate *Nr*, which is why the connectivity of the GHZ state is *N*. By contrast, for RVB states, there are *N*/2 singlets, so half of the qubits need to be removed to destroy entanglement. Similarly, this state is not markedly susceptible to dephasing, indicating a low value of connectivity.

Given that macroscopic entanglement exists, an important technological question is how easy this entanglement would be to extract and use. Suppose that two neutron beams are aimed at a magnetic substance, each at a different section[31]. It is fruitful — albeit not entirely mathematically precise — to think of this interaction as a state swap of the spins of the neutrons and the spins of the atoms in the solid. If the atoms in the solid are themselves entangled, then this entanglement is transferred to each of the scattered spins. This transferral could then presumably be used for further information processing. Similarly, schemes can be designed to extract entanglement from Bose–Einstein condensates[32,33] and superconductors[34] (which can be thought of as Bose–Einstein condensates of Cooper pairs of electrons), although none of these extraction schemes has been implemented as yet.

There are many open questions regarding entanglement. Here I have stated that, in theory, entanglement can exist in arbitrarily large and hot systems. But how true is this in practice? Another question is whether the entanglement of massless bodies fundamentally differs from that of massive ones[35]. Furthermore, does macroscopic entanglement also occur in living systems and, if so, is it used by these systems?

Some of the open questions might never be answered. Some might turn out to be uninteresting or irrelevant. One thing is certain though: current experimental progress is so rapid that future findings will surprise researchers and will take the current knowledge of entanglement to another level. ■

1. Amico, L., Fazio, R., Osterloh, A. & Vedral, V. Many-body entanglement. *Rev. Mod. Phys.* **80,** 517–576 (2008).
2. Osterloh, A. *et al.* Scaling of entanglement close to a quantum phase transition. *Nature* **416,** 608–610 (2002).
3. Osborne, T. J. & Nielsen, M. A. Entanglement in a simple quantum phase transition. *Phys. Rev. A* **66,** 032110 (2002).
4. Arnesen, M. C., Bose, S. & Vedral, V. Natural thermal and magnetic entanglement in 1D Heisenberg model. *Phys. Rev. Lett.* **87,** 017901 (2001).
5. Vedral, V. High temperature macroscopic entanglement. *New J. Phys.* **6,** 102 (2004).
6. Brukner, C., Vedral, V. & Zeilinger, A. Crucial role of entanglement in bulk properties of solids. *Phys. Rev. A* **73,** 012110 (2006).
7. Vedral, V. A better than perfect match. *Nature* **439,** 397 (2006).
8. Zeilinger, A., Weihs, G., Jennewein, T. & Aspelmeyer, M. Happy centenary, photon. *Nature* **433,** 230–238 (2005).
9. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47,** 777–780 (1935).
10. Werner, R. F. Quantum states with Einstein–Podolsky–Rosen correlations admitting a hidden-variable model. *Phys. Rev. A* **40,** 4277–4281 (1989).
11. Brukner, C. & Vedral, V. Macroscopic thermodynamical witnesses of quantum entanglement. Preprint at <http://arxiv.org/abs/quant-ph/0406040> (2004).
12. Toth, G. & Guhne, O. Detecting genuine multipartite entanglement with two local measurements. *Phys. Rev. Lett.* **94,** 060501 (2004).
13. Narnhoffer, H. Separability for lattice systems at high temperature. *Phys. Rev. A* **71,** 052326 (2005).
14. Schrödinger, E. Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften* **23,** 807–812; 823–828; 844–849 (1935).
15. Horodecki, M., Horodecki, P. & Horodecki, R. Separability of mixed states: necessary and sufficient conditions. *Phys. Lett. A* **223,** 1–8 (1996).
16. Anders J. & Vedral, V. Macroscopic entanglement and phase transitions. *Open Sys. Inform. Dyn.* **14,** 1–16 (2007).
17. Horodecki, M. Entanglement measures. *Quant. Inform. Comput.* **1,** 3–26 (2001).
18. Vedral V. *et al.* Quantifying entanglement. *Phys. Rev. Lett.* **78,** 2275–2279 (1997).
19. Greenberger, D., Horne, M. A. & Zeilinger, A. in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe* (ed. Kafatos, M.) 73–76 (Kluwer Academic, Dordrecht, 1989).
20. Dur, W., Vidal, G. & Cirac, J. I. Three qubits can be entangled in two inequivalent ways. *Phys. Rev. A* **62,** 062314 (2000).
21. Leggett, A. J. Macroscopic quantum systems and the quantum theory of measurement. *Prog. Theor. Phys. Suppl.* **69,** 80–100 (1980).
22. Anders, J. & Winter, A. Entanglement and separability of quantum harmonic oscillator systems at finite temperature. *Quant. Inform. Comput.* **8,** 0245–0262 (2008).
23. Vedral, V. Entanglement in the second quantisation formalism. *Cent. Eur. J. Phys.* **2,** 289–306 (2003).
24. Anderson, P. W. Resonating valence bonds: a new kind of insulator? *Mater. Res. Bull.* **81,** 53–60 (1973).
25. Raussendorf, R. & Briegel, H. J. A one-way quantum computer. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
26. Chandran, A., Kaszlikowski, D., Sen De, A., Sen, U. & Vedral, V. Regional versus global entanglement in resonating-valence-bond states. *Phys. Rev. Lett.* **99,** 170502 (2007).
27. Page, D. N. & Wootters, W. K. Evolution without evolution: dynamics described by stationary observables. *Phys. Rev. D* **27,** 2885–2892 (1983).
28. Haffner, H. *et al.* Scalable multiparticle entanglement of trapped ions. *Nature* **438,** 643–646 (2005).
29. Lu, C.-Y. *et al.* Experimental entanglement of six photons in graph states. *Nature Phys.* **3,** 91–95 (2007).
30. Baugh, J. *et al.* Quantum information processing using nuclear and electron magnetic resonance: review and prospects. Preprint at <http://arxiv.org/abs/0710.1447> (2007).
31. de Chiara, G. *et al.* A scheme for entanglement extraction from a solid. *New J. Phys.* **8,** 95 (2006).
32. Toth, G. Entanglement detection in optical lattices of bosonic atoms with collective measurements. *Phys. Rev. A* **69,** 052327 (2004).
33. Heaney, L., Anders, J., Kaszlikowski, D. & Vedral, V. Spatial entanglement from off-diagonal long-range order in a Bose–Einstein condensate. *Phys. Rev. A* **76,** 053605 (2007).
34. Recher, P. & Loss, D. Superconductor coupled to two Luttinger liquids as an entangler for spin electrons. *Phys. Rev. B* **65,** 165327 (2002).
35. Verstraete, F. & Cirac, J. I. Quantum nonlocality in the presence of superselection rules and data hiding protocols. *Phys. Rev. Lett.* **91,** 010404 (2003).

Acknowledgements I am grateful for funding from the Engineering and Physical Sciences Research Council, the Wolfson Foundation, the Royal Society and the European Union. My work is also supported by the National Research Foundation (Singapore) and the Ministry of Education (Singapore). I thank J. A. Dunningham, A. J. Leggett, D. Markham, E. Rieper, W. Son and M. Williamson for discussions of this and related subjects. W. Son's help with illustrations is also gratefully acknowledged.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (vlatko.vedral@quantuminfo.org).

NATURE|Vol 453|19 June 2008|doi:10.1038/nature07125

Entangled states of trapped atomic ions

Rainer Blatt[1,2] & David Wineland[3]

To process information using quantum-mechanical principles, the states of individual particles need to be entangled and manipulated. One way to do this is to use trapped, laser-cooled atomic ions. Attaining a general-purpose quantum computer is, however, a distant goal, but recent experiments show that just a few entangled trapped ions can be used to improve the precision of measurements. If the entanglement in such systems can be scaled up to larger numbers of ions, simulations that are intractable on a classical computer might become possible.

For more than five decades, quantum superposition states that are coherent have been studied and used in applications such as photon interferometry and Ramsey spectroscopy[1]. However, entangled states, particularly those that have been 'engineered' or created for specific tasks, have become routinely available only in the past two decades (see page 1004). The initial experiments with pairs of entangled photons[2,3], starting in the 1970s, were important because they provided tests of non-locality in quantum mechanics[4]. Then, in the early to mid-1980s, Richard Feynman and David Deutsch proposed that it might be possible way to carry out certain computations or quantum simulations efficiently by using quantum systems[5,6]. This idea was, however, largely considered a curiosity until the mid-1990s, when Peter Shor devised an algorithm[7] that could factor large numbers very efficiently with a quantum computer. This marked the beginning of widespread interest in quantum information processing[8] and stimulated several proposals for the implementation of a quantum computer.

Among these proposals, the use of trapped ions[9] has proved to be one of the most successful ways of deterministically creating entangled states, and for manipulating, characterizing and using these states for measurement. At present, about 25 laboratories worldwide are studying aspects of quantum information processing with trapped ions. Ions provide a relatively 'clean' system, because they can be confined for long durations while experiencing only small perturbations from the environment, and can be coherently manipulated. Although trapping ions in this way involves several technical processes, the system is an accessible one in which to test concepts that might be applicable to other systems, such as those involving neutral trapped atoms, quantum dots, nuclear spins, Josephson junctions or photons.

In this review, we highlight recent progress in creating and manipulating entangled states of ions, and we describe how these advances could help to generate quantum gates for quantum information processing and improve tools for high-precision measurement. For a review of earlier progress in quantum information processing with atoms, including atomic ions, and photons, see ref. 10.

Trapped and laser-cooled ions

To study entanglement, it is desirable to have a collection of quantum systems that can be individually manipulated, their states entangled, and their coherences maintained for long durations, while suppressing the detrimental effects of unwanted couplings to the environment. This can be realized by confining and laser cooling a group of atomic ions in a particular arrangement of electric and/or magnetic fields[11,12]. With such 'traps', atomic ions can be stored nearly indefinitely and can be localized in space to within a few nanometres. Coherence times of as long as ten minutes have been observed for superpositions of two hyperfine atomic states of laser-cooled, trapped atomic ions[13,14].

In the context of quantum information processing, a typical experiment involves trapping a few ions by using a combination of static and sinusoidally oscillating electric potentials that are applied between the electrodes of a linear quadrupole, an arrangement known as a Paul trap[12] (Fig. 1). When the trapped ions are laser cooled, they form a linear 'string', in which the spacings are determined by a balance between the horizontal (axial) confining fields and mutual Coulomb repulsion. Scattered fluorescence, induced by a laser beam, can be imaged with a camera (Fig. 1). The use of tightly focused laser beams allows the manipulation of individual ions.

For simplicity, in this review, we focus on two specific internal states of each ion, which we refer to as the ground and excited states ($|g\rangle$ and $|e\rangle$, respectively). This 'quantum bit' (qubit) structure is 'dressed' by the oscillator states $|n\rangle$ of frequency ω_m of a particular mode (Fig. 1). We denote the internal states as 'spin' states, in analogy to the two states of a spin −½ particle. If the energy between internal states corresponds to an optical frequency ω_{eg}, this atomic transition can be driven by laser radiation at frequency ω_{eg}, which couples states $|g, n\rangle \leftrightarrow |e, n\rangle$, where $|g, n\rangle$ denotes the combined state $|g\rangle|n\rangle$. Spin and motional degrees of freedom can be coupled by tuning the laser to 'sideband' frequencies $\omega_{eg} \pm \omega_m$, which drives transitions $|g, n\rangle \leftrightarrow |e, n + \Delta n\rangle$ (refs 15–18), with $\Delta n = \pm 1$. In this case, state evolution can be described as a rotation $R_{\Delta n}(\theta, \phi)$ of the state vector on the Bloch sphere[8,18] and is defined here as

$$R_{\Delta n}(\theta, \phi)\,|g, n\rangle \rightarrow \cos\frac{\theta}{2}|g, n\rangle + ie^{i\phi}\sin\frac{\theta}{2}|e, n+\Delta n\rangle$$

$$R_{\Delta n}(\theta, \phi)\,|e, n+\Delta n\rangle \rightarrow ie^{-i\phi}\sin\frac{\theta}{2}|g, n\rangle + \cos\frac{\theta}{2}|e, n+\Delta n\rangle \quad (1)$$

where θ depends on the strength and the duration of the applied laser pulse, ϕ is the laser beam phase at the ion's position and $i = \sqrt{-1}$. For $\Delta n = \pm 1$, entanglement is generated between the spin and motional degrees of freedom. Higher-order couplings ($|\Delta n| > 1$) are suppressed for laser-cooled ions, the spatial extent of which is much smaller than the laser wavelength, which is known as the Lamb–Dicke regime. In this regime, sideband laser cooling works by tuning the laser to induce absorption on the lower sideband frequency ($\Delta n = -1$), followed by spontaneous emission decay, which occurs mainly at the 'carrier'

[1]Institut für Experimentalphysik, Universität Innsbruck, Technikerstrasse 25, A-6020 Innsbruck, Austria. [2]Institut für Quantenoptik und Quanteninformation, Österreichische Akademie der Wissenschaften, Otto-Hittmair-Platz 1, A-6020 Innsbruck, Austria. [3]National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305, USA.

transition frequency ($\Delta n = 0$). With repeated absorption–emission cycles, the ions are optically pumped to the combined spin and motion ground state $|g, n=0\rangle$ (ref. 19). If the spin energy levels correspond to microwave or lower frequencies (as occurs in hyperfine atomic states and Zeeman states), the same processes can be realized by replacing single-photon optical transitions with two-photon stimulated-Raman transitions and by replacing spontaneous emission with spontaneous Raman scattering[15–18]. It should be noted that there are similarities between the coupling of an ion's internal states to the harmonic oscillator associated with a mode of motion and the case of cavity quantum electrodynamics, in which an atom's internal states are coupled to the harmonic oscillator associated with a single electromagnetic mode of the cavity (see page 1023).

The qubit state of an ion can be detected with more than 99% efficiency by measuring resonance fluorescence from an auxiliary state that is strongly coupled (by a monitoring excitation) to one of the qubit states ($|g\rangle$ or $|e\rangle$) and decays back only to that same state, known as a cycling transition. This is usually called quantum non-demolition (QND) detection because when the ion has been projected into a particular spin state, it will remain in that state throughout repeated excitation–emission cycles. Therefore, a cycle can be repeated many times, and it is not necessary to detect every emitted photon to obtain a high overall detection efficiency. If the qubit is projected to, or 'shelved' in, the state that is not coupled to the fluorescing transition, then no photons are observed, and this state can therefore be distinguished from the fluorescing state[20].

Spin-entangled states

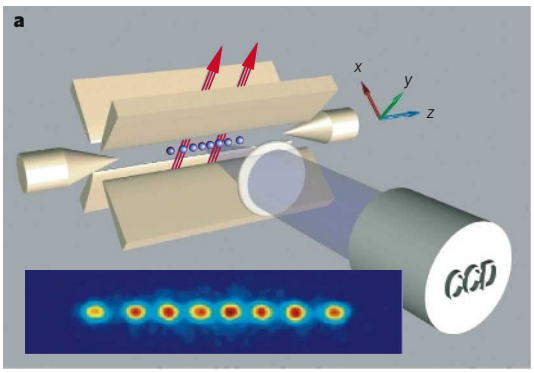
In 1995, Ignacio Cirac and Peter Zoller suggested how to use a trapped-ion system to implement a quantum computer[9]. For universal quantum computing and for the generation of arbitrary entangled qubit states, two basic gate operations are required: first, individual qubit rotations as described by equation (1); and, second, a two-qubit-entangling operation that is the quantum counterpart to the classical operation with the XOR logic gate, the controlled-NOT (CNOT)-gate operation. The CNOT gate flips the state of a target qubit depending on the state of a control qubit. And, importantly, when applied to superposition states, it generates entanglement. The CNOT operation (Fig. 2) is achieved with a sequence of carrier pulses ($R_0(\theta, \phi)$) and red sideband pulses ($R_{-1}(\theta, \phi)$). The central part of this sequence involves a 'phase gate' that applies a phase shift $e^{i\pi} = -1$ to the $|g, n=1\rangle$ component of the target ion's wavefunction. This is implemented by applying a coherent $R_{-1}(2\pi, \phi)$ pulse between the $|g, 1\rangle$ state and an auxiliary state $|\text{aux}, 0\rangle$. Because the applied radiation cannot excite the states $|g, 0\rangle$, $|e, 0\rangle$ or $|e, 1\rangle$, they are unaffected. This operation is sandwiched between rotations that transfer phase changes into state changes, as occurs in Ramsey spectroscopy. By using a single ion, Christopher Monroe *et al.*[21] realized the CNOT operation between motion and spin for $^9Be^+$ ions. Subsequently, Ferdinand Schmidt-Kaler *et al.*[22,23] and later Mark Riebe *et al.*[24] realized the complete CNOT operation between two individually addressed $^{40}Ca^+$ ions. Entangling gates have also been realized by irradiating ions simultaneously (Fig. 3). Although such gates can be implemented in a single step, they still involve transitory entanglement with a motional mode, which effectively couples the spin qubits. Ions have also been entangled with each other in a probabilistic way mediated by entanglement with scattered photons[25] (Fig. 4).

By sequentially combining single-qubit and multiqubit operations, various entangled states of ions have been created deterministically or 'on demand'. A research group from the National Institute of Standards and Technology (NIST), in Boulder, Colorado, created[26] the state $|\Psi_e(\phi)\rangle = \frac{3}{5}|ge\rangle - e^{i\phi}\frac{4}{5}|eg\rangle$, where ϕ is a controllable phase factor and $|ge\rangle$ denotes the combined state $|g\rangle_1|e\rangle_2$ for ions 1 and 2. More generally, by using entangling operations and single-qubit rotations with adjustable phases, all Bell states — $|\Psi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|ge\rangle \pm |eg\rangle)$, $|\Phi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|gg\rangle \pm |ee\rangle)$ — and arbitrary superpositions can be generated[27,28]. The quality or fidelity of quantum states is usually characterized by the degree with which they agree with the desired (or ideal) state, which is expressed as

$$F = \langle \Psi_{\text{ideal}} | \rho_{\text{exp}} | \Psi_{\text{ideal}} \rangle \quad (2)$$

where ρ_{exp} is the experimentally achieved density matrix, which characterizes both pure and non-pure states. In current experiments, fidelities $F > 0.95$ are achieved.

In some cases, complete knowledge of the density matrix is not required. For example, the fidelity of a state relative to $|\Phi^+\rangle$ can be derived from just three matrix elements, $F = \frac{1}{2}(\rho_{gg,gg} + \rho_{ee,ee}) + \text{Re}\rho_{ee,gg}$, where $\rho_{ee,gg} \equiv \langle ee|\rho_{\text{exp}}|gg\rangle$ and so on and Re denotes the real part of the expression that follows. The matrix elements $\rho_{gg,gg}$ and $\rho_{ee,ee}$ are obtained from the measured populations of the respective states. The matrix element $\rho_{ee,gg}$



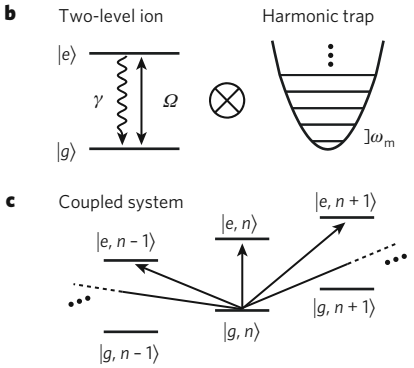


Figure 1 | Ions confined in a trap. a, A linear quadrupole ion trap (known as a Paul trap; beige) containing individually addressed $^{40}Ca^+$ ions (blue) is depicted. After cooling by laser beams (red), the trapped ions form a string and are then imaged by using a charge-coupled device (CCD). In the CCD image shown, the spacing of the two centre ions is ~8 μm. The electrode arrangement in the Paul trap provides an almost harmonic three-dimensional well. For a single ion, this is characterized by three frequencies[17]: ω_x, ω_y and ω_z, where x, y and z denote the confining potential axes. In this case, z points along the trap axis and x, y in the transverse directions. Owing to the Coulomb coupling that occurs between ions, the motion is best described in terms of normal modes; a string of ions can therefore be viewed as a pseudo-molecule. In general, the normal-mode frequencies ω_m differ from each other, and a particular mode can be accessed by spectral selection. **b**, The energy levels of a two-level ion (left) and one mode of the ion's motion (right) are shown. On the left is depicted the ion's ground state $|g\rangle$ and excited state $|e\rangle$, interacting with radiation characterized by the Rabi frequency Ω and decaying with the rate γ. On the right is depicted the harmonic oscillator potential and equally spaced energy levels for one mode of motion. Both the two-level system and the harmonic oscillator can be described jointly in a quantum-mechanical way, indicated by the direct product $\otimes$, resulting in a manifold of two-level systems separated by the mode frequency ω_m (as shown in **c**). **c**, The level structure of the coupled ion–harmonic-oscillator system is shown, with states jointly described by the spin ($|g\rangle$ and $|e\rangle$) and motional ($|0\rangle, |1\rangle, \ldots, |n\rangle$) degrees of freedom, where $|g\rangle|n\rangle = |g, n\rangle$ and $|e\rangle|n\rangle = |e, n\rangle$. Arrows indicate the transitions that are possible when appropriately tuned radiation is applied; dashed lines indicated connections to levels not shown.

can be obtained by applying a rotation $R_0(\pi/2, \phi)$ to both ions and measuring the parity $P \equiv |gg\rangle\langle gg| + |ee\rangle\langle ee| - |ge\rangle\langle ge| - |eg\rangle\langle eg|$ of the resultant state as a function of ϕ. The only component of the parity that oscillates sinusoidally with frequency 2ϕ is proportional to $\rho_{ee,gg}$, which allows this element to be extracted[29].

As shown by equation (2), the fidelity can be obtained by measuring the full density matrix. To do this, the quantum state in question must be prepared many times; in this way, with the appropriate single-qubit rotations applied before the qubit measurements, all expectation values of the density matrix are obtained. Such a procedure is known as quantum-state tomography[28]. When this procedure is applied to Bell states, the density matrix can be completely characterized (Fig. 5). From the density matrices, all measures can subsequently be calculated. For example, in the case of Bell's inequalities, it is possible to determine the expectation value of the operator[30] $A = \sigma_x^{(1)} \otimes \sigma_{x-z}^{(2)} + \sigma_x^{(1)} \otimes \sigma_{x+z}^{(2)} + \sigma_z^{(1)} \otimes \sigma_{x-z}^{(2)} - \sigma_z^{(1)} \otimes \sigma_{x+z}^{(2)}$, where $\sigma_{x\pm z} = (\sigma_x \pm \sigma_z)/\sqrt{2}$ and σ is a Pauli operator and the superscripts refer to the first and second qubits. For local realistic theories, measurements of $|\langle A \rangle|$ are predicted to be less than 2, and values of $2 < |\langle A \rangle| < 2\sqrt{2}$ are expected for states that can be described only by quantum theory. With trapped ions, experiments yielded $|\langle A \rangle| = 2.25(3)$ at NIST[27], $|\langle A \rangle| = 2.52(6)$ at the Institute for Experimental Physics, University of Innsbruck (Innsbruck, Austria)[28], and $|\langle A \rangle| = 2.20(3)$ at the FOCUS Center and Department of Physics, University of Michigan (Ann Arbor, Michigan)[31], where the number in parentheses denotes the uncertainty in the last digit, clearly corroborating quantum theory (Fig. 5). Moreover, each time an experiment was run, a result was recorded. This closed the 'detection loophole', which would provide a way to violate Bell's inequalities within local realistic theories.

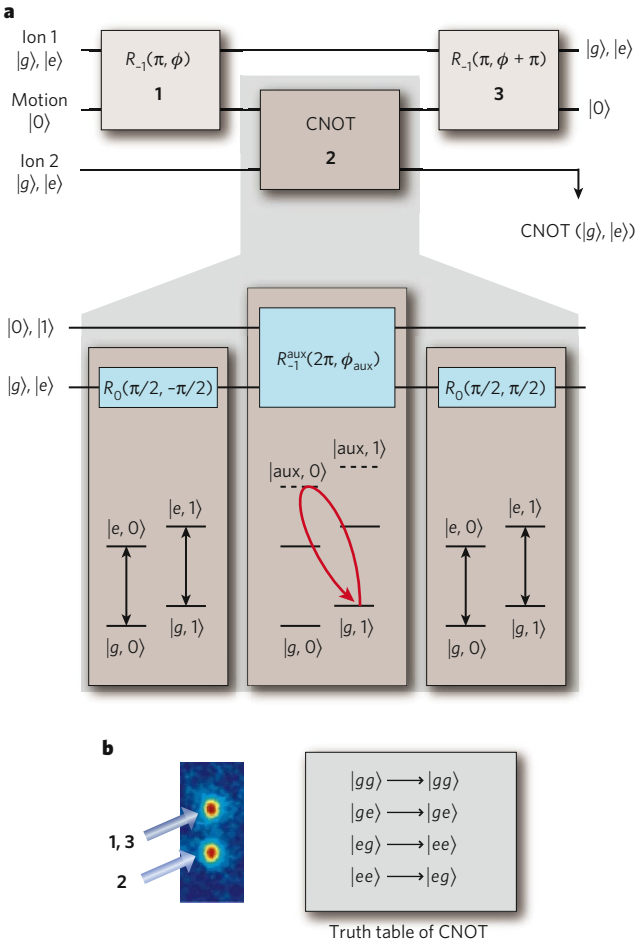


Figure 2 | A CNOT-gate operation with two trapped ions. a, Consider two ions in the same trap that are initially prepared in their motional ground state. In step 1, a lower-sideband pulse $R_{-1}(\pi, \phi)$ is applied to the first ion (ion 1; the control qubit) and maps the excited-state amplitude of this ion to the first excited state of a selected motional mode (a process known as a SWAP operation). Importantly, this motional excitation is also shared with the second ion (ion 2; the target qubit). In step 2, a CNOT-gate operation is implemented between the motion qubit (which is shared by both spin qubits) and the spin state of ion 2. Finally, in step 3, the first step is reversed, thereby restoring the initial spin state of ion 1 and returning the motion to its ground state. The pulse sequence of the CNOT-gate operation is also shown (lower part of **a**). **b**, On the left is a CCD image of two ions. The arrows indicate laser radiation that is applied to the ions in the order of the indicated numbers (which correspond to the three steps in **a**). First, a laser pulse is applied to the upper ion (1), then the CNOT sequence is applied to the lower ion (2). Finally, a laser pulse is applied to the upper ion again (3). On the right is the resultant truth table of the CNOT-gate operation, with the first and second symbols denoting the state of the control qubit (ion 1) and the target qubit (ion 2), respectively.

The operations outlined above can be generalized to entangle more than two particles. Among such states, the 'cat' states, named after Schrödinger's cat[32], are of particular interest. Cat states are usually defined as superpositions of two particular maximally different states, such as $|\Psi_{cat}\rangle = \alpha|ggg\ldots g\rangle + \beta|eee\ldots e\rangle$, and they have an important role in quantum information science. For three qubits, cat states are also known as GHZ states, which were named after Daniel Greenberger, Michael Horne and Anton Zeilinger, who showed that these states could provide a particularly clear contradiction with local realistic theories[33]. They are a fundamental resource in fault-tolerant quantum computing, for error correction[34,35] and for quantum communication. In addition, because of their sensitivity to the interferometric phase ϕ, they can also improve signal-to-noise ratios in interferometry[36] (described later).

With trapped ions, cat states with $|\alpha| = |\beta|$ have been generated by using two approaches. At NIST, global entangling operations were used to demonstrate a cat state of four ions[29], a GHZ state with $F = 0.89$ (ref. 37), and cat states of up to six ions[38]. Using individually addressed ions and a CNOT-gate operation, the research group at Innsbruck produced GHZ states in an algorithmic way and analysed the states by using tomographic measurements[39]. In a similar way, the Innsbruck group also produced W states for N-ion qubits, $|\Psi_W\rangle = \frac{1}{\sqrt{N}}(|g\ldots gge\rangle + |g\ldots geg\rangle + \ldots + |eg\ldots g\rangle)$, which belong to a different class of entangled states. Such classes are distinct because states of different classes cannot be transformed into each other by local operations and classical communication[40]. Nevertheless, both cat and W states can violate Bell-type inequalities. In contrast to cat states, W states are remarkably robust in the face of a variety of decoherence processes: for W states, even the loss of qubits does not destroy entanglement completely. The Innsbruck group deterministically prepared an eight-qubit W state[41] by using individual ion addressing. Both the NIST and Innsbruck groups verified multipartite entanglement by using an 'entanglement witness', an operator constructed so that its expectation value must exceed (or be less than) a certain value to verify N-particle entanglement[38,41].

Demonstrating quantum-information-processing algorithms

Algorithms are lists of instructions for completing a task[8]. As is the case in classical computation, quantum algorithms can sometimes be viewed as subroutines in a larger computation. A quantum-information-processing algorithm generally involves single-qubit gates and multiqubit gates, as well as measurements and measurement-dependent operations. The result of such a sequence of operations could be a deterministically prepared quantum state (such as a Bell, GHZ or W state), a conditioned state (such as an error-corrected state) or a state that is subsequently inferred from a measurement of the quantum register and is then available as classical information.

In contrast to classical information processing, quantum information processing allows tests to be carried out using superpositions. A simple example showing the gain in efficiency that is possible with a quantum algorithm was proposed by Deutsch and Richard Jozsa[42]. The Deutsch–Jozsa algorithm was first demonstrated with two qubits in nuclear magnetic resonance spectroscopy[43], and it was demonstrated more recently with a trapped ion[44], with the motional and spin properties of the ion qubit serving as the two qubits.

Another example algorithm is teleportation of the state of one qubit to another qubit, an important protocol for the transfer of quantum information[10,45]. In this algorithm, Alice wants to send a qubit state (which, in general, is unknown) to Bob. To do this, a Bell pair is generated, and one qubit from this pair is given to the sender, Alice, and the other qubit to the receiver, Bob. When the unknown state is ready to be teleported, it is entangled with Alice's qubit of the Bell pair. A subsequent measurement of both qubits by Alice yields two bits of classical information that she sends to Bob. With this information, Bob knows which of four possible rotations to apply to his qubit to obtain Alice's original unknown state.

Deterministic quantum teleportation has been demonstrated by the NIST[46] and Innsbruck[47] groups. The Innsbruck group used individual laser-beam addressing of three qubits; therefore, the state was teleported from one end of the ion string to the other end, a distance of ~10 μm. The NIST group used a multizone linear-trap array. By applying control potentials to electrode segments, the ions could be separated and moved in and out of one zone in which the laser beams were present. In this case, the state was teleported across a few hundred micrometres.

Teleportation is an important building block for quantum information processing and can reduce the computational resource requirements. Furthermore, it is the basic procedure for quantum communication protocols, such as for implementing quantum repeaters. Other algorithms — such as entanglement purification[48], quantum error correction[49], the quantum Fourier transform[50] and deterministic entanglement swapping (M. Riebe, T. Monz, K. Kim, A. S. Villar, P. Schindler, M. Chwalla, M. Hennrich and R. Blatt, unpublished observations) — have also been demonstrated with ion qubits.

These experiments demonstrate the basic features of quantum algorithms, but for the concatenation of processes and repeated computations, improved operation fidelities will be required. In particular, full and repetitive implementation of quantum error correction, which could keep a qubit superposition 'alive' while subjected to decoherence, remains a major challenge in quantum information processing.

Applications

In the mid-1990s, a wave of interest in potential applications for quantum information processing was generated by Shor's period-finding algorithm for factoring large numbers[7]. Another noteworthy potential application is the implementation of unstructured searches[51]. However, to be of practical use, these applications require substantial resources in terms of the number of qubits and the number of operations, far beyond the capabilities of current implementations. Despite this, some elements of quantum information processing and entanglement with small numbers of qubits are beginning to find applications in metrology. Many physicists also expect that useful quantum simulations will be carried out on a relatively small number of qubits, perhaps up to 100, in the next decade.

One application in metrology is to improve interferometry. As an example, we discuss how entanglement can be applied to Ramsey spectroscopy[52], but this scheme has a direct analogue in electron, atom and photon Mach–Zehnder interferometry. Ramsey spectroscopy on the $|g\rangle \rightarrow |e\rangle$ transition proceeds as follows. The atom is first prepared in the state $|\Psi_{\text{initial}}\rangle = |g\rangle$. Radiation at frequency ω near ω_{eg} is applied in a fast pulse to produce the state $R_0(\pi/2, -\pi/2)|g\rangle = \frac{1}{\sqrt{2}}(|g\rangle + |e\rangle)$. The atom is now allowed to evolve for a duration T so that the atom's upper state accumulates a phase $\phi_R = (\omega - \omega_{ge})T$ relative to the lower state (when the problem is viewed in a frame that rotates at frequency ω). Finally, again, a rotation $R_0(\pi/2, -\pi/2)$ is applied and leaves the atom in the state (up to a global phase factor) $|\Psi_{\text{final}}\rangle = \sin(\phi_R/2)|g\rangle + i\cos(\phi_R/2)|e\rangle$. Therefore, the probability of finding the atom in the state $|e\rangle$ is $p_e = \frac{1}{2}(1 + \cos[(\omega - \omega_{ge})T])$. For an ensemble of N atoms, the detected signal will be Np_e. In precision spectroscopy, the idea is to detect changes in $\omega - \omega_{ge}$ or ϕ_R, as observed through changes in p_e. Therefore, the N-ion signal can be defined as $S = \mathrm{d}(Np_e)/\mathrm{d}\phi_R = -N/2\sin(\phi_R)$. The fundamental noise in the signal is given by the 'projection noise': that is, the fluctuation in the number of atoms, from experiment to experiment, that is measured to be in the state $|e\rangle$ (ref. 53). The variance of this noise is given by $V_N = Np_e(1 - p_e)$, so the magnitude of the signal-to-noise ratio is equal to $S/\sqrt{V_N} = \sqrt{N}$, essentially the shot noise corresponding to the number of atoms.

Now, suppose that the first $R_0(\pi/2, -\pi/2)$ pulse can be replaced with an entangling $\pi/2$ pulse[37,38], which creates the cat state

$$|g\rangle_1|g\rangle_2 \ldots |g\rangle_N \rightarrow \tfrac{1}{\sqrt{2}}(|g\rangle_1|g\rangle_2 \ldots |g\rangle_N + |e\rangle_1|e\rangle_2 \ldots |e\rangle_N) \equiv \tfrac{1}{\sqrt{2}}(|g_N\rangle + |e_N\rangle) \qquad (3)$$

After a delay T, the $|e_N\rangle$ state accumulates a phase $N\phi_R$ relative to the $|g_N\rangle$ state. A final entangling $\pi/2$ pulse leaves the atoms in a superposition state $\sin(N\phi_R/2)|g_N\rangle + i\cos(N\phi_R/2)|e_N\rangle$; therefore, the probability of detecting the atoms in the $|e_N\rangle$ state is $p_{N_e} = \frac{1}{2}(1 + \cos[N(\omega - \omega_{ge})T])$. It is as though spectroscopy has been carried out on a single 'super-atom' composed of states $|e_N\rangle$ and $|g_N\rangle$. The super-atom has a resonant frequency that is N times higher than that of a single atom, as well as a phase sensitivity (to the Nth harmonic of the applied radiation) that is N times higher. The resultant gain in interferometric sensitivity must, however, be offset by the fact that only a single-particle two-state system ($|e_N\rangle$ and $|g_N\rangle$) is being measured. Nevertheless, after a statistically significant number of repeated measurements, the sensitivity is

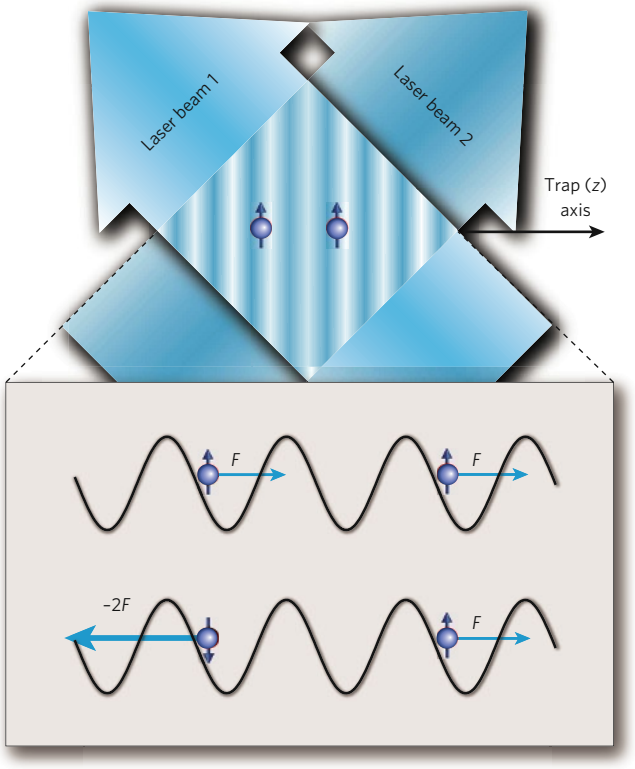


Figure 3 | A two-qubit phase gate. A phase gate with two ions (blue) is depicted. The operation of such phase gates relies on the fact that when a selected mode of the ions' motion is displaced in phase space about a closed path, the ions' wavefunction picks up a phase that is proportional to the enclosed area. If this displacement depends on the ions' qubit states, then entanglement is generated[92–95]. This state-dependent displacement can be implemented by applying optical dipole forces (F) by using laser-beam intensity gradients. In this example, an intensity standing wave is created with two laser beams, and the horizontal spacing of the ions is made to be an integral number of wavelengths of the intensity pattern. The pattern sweeps across the ions at the difference between the frequencies of the beams, chosen to be near the stretch-mode frequency. If the ions' qubit states $|g\rangle$ and $|e\rangle$ feel different dipole forces, then only the $|ge\rangle$ and $|eg\rangle$ components of the ions' wavefunction are displaced in phase space. By making the trajectories closed and by choosing the size of the displacements appropriately, the wavefunction is unchanged except for an $e^{i\pi/2}$ phase shift on the $|ge\rangle$ and $|eg\rangle$ states, the desired phase gate. Such gate operations have been implemented with trapped $^9Be^+$ ions[29,95] and in a similar way with $^{111}Cd^+$ ions[96] and $^{40}Ca^+$ ions[63,97].

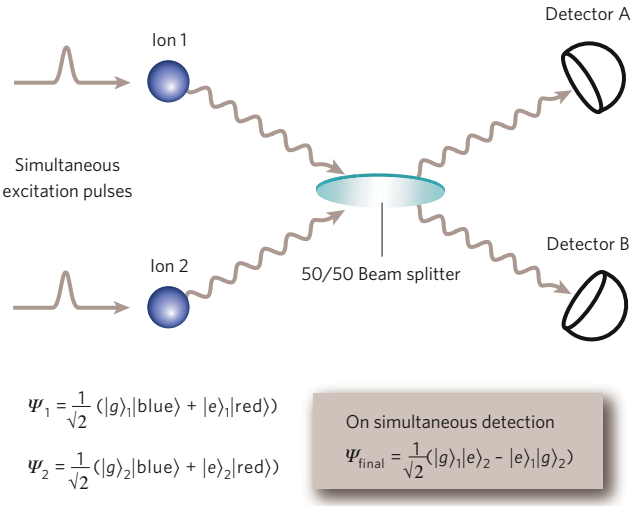


Figure 4 | Entanglement produced by conditional measurements. Entanglement can be created between two separated particles by an interference effect and state projection accompanying a measurement. In this example[25], it is assumed that the qubits of two ions (blue) are encoded in hyperfine levels of the electronic ground states. These qubits are first prepared in superposition states $\frac{1}{\sqrt{2}}(|g\rangle + |e\rangle)$. When excited with laser pulses that are short enough that both qubits simultaneously undergo (single-photon) scattering, the frequencies (denoted 'red' and 'blue') of the emitted photons along a particular direction are correlated with the qubit states, as indicated for entangled states $|\Psi_1\rangle$ and $|\Psi_2\rangle$. These photons can be simultaneously sent into a 50/50 beam splitter and then detected. In the cases when photons are simultaneously detected at detector A and detector B, the ions are projected into the Bell state $|\Psi_{\text{final}}\rangle$, even though the atoms have not directly interacted. For many such experiments, photons do not reach either detector; however, when photons are simultaneously detected, this 'heralds' the formation of the entangled state $|\Psi_{\text{final}}\rangle$, which can then be saved and used later, such as in Bell's inequality measurements of remotely located ions[98]. One potential use of this scheme is for entanglement-assisted communication between ion locations 1 and 2.

increased by a factor of $\sqrt{N}$ by using entangling $\pi/2$ pulses compared with the case of N unentangled atoms[36–38]. Because of technical noise in the experiments, this theoretical improvement is not fully realized; however, a gain in sensitivity compared with the case of unentangled atoms has been realized for up to six entangled ions[38,54,55].

These arguments assume that noise results only from state projection. In experiments, if there is correlated decoherence of qubit phases, then any gain in sensitivity may be lost as a result of the faster decoherence of the cat states[56] or as a result of noise in the oscillator that produces the radiation[18,57]. If these sources of noise can be suppressed, entangled states should be able to improve the signal-to-noise ratio in future spectroscopy experiments.

Another application of quantum-information-processing techniques is increased fidelity of detection[58]. This can be useful if the qubit does not have a cycling transition or if the QND aspect of the shelving detection is not well satisfied. A simple implementation is to assume that there are two qubits, labelled q and d, stored in the same trap. The goal is to detect the state of information qubit q, by using detection qubit d. Before any measurements are taken, qubit q will generally be in a superposition state $\alpha|g\rangle_q + \beta|e\rangle_q$. Using the SWAP operations of the Cirac–Zoller gate, this superposition is first transferred to the qubit composed of the $|0\rangle$ and $|1\rangle$ states of a selected motional mode, and is then mapped to qubit d. Then, qubit d is measured, thereby in effect measuring qubit q. This protocol can be carried out without disturbing the initial probabilities $|\alpha|^2$ and $|\beta|^2$ for qubit q, even if the mapping steps are imperfect. Therefore, it is a QND measurement and can be repeated to increase detection efficiency. This scheme was demonstrated in an experiment[59] in which qubit q was based on an optical transition in a $^{27}Al^+$ ion and qubit d was based on a hyperfine transition in a $^9Be^+$ ion. In that experiment, a single round of detection had a fidelity of only $F = 0.85$; however, by repeating the measurement, and by using real-time bayesian analysis, the fidelity was improved to $F = 0.9994$. It should be noted that this strategy can also be used to prepare an eigenstate of qubit q with high fidelity. In addition to this demonstration, this protocol is now used in a high-accuracy optical clock based on single $^{27}Al^+$ ions[60]. This technique has also been extended so that a single detection qubit can be used to measure the states of multiple ions[59], similar to the measurement of the Fock states of photons by using multiple probe atoms[61].

Finally, entanglement can be used in metrology to create states that allow the measurement of certain parameters while suppressing sensitivity to others. This strategy has been used, for example, to make a precise measurement of the quadrupole moment of a $^{40}Ca^+$ ion by carrying out spectroscopy on an entangled state of two ions that depended on the quadrupole moment but was insensitive to fluctuations in the magnetic field[62].

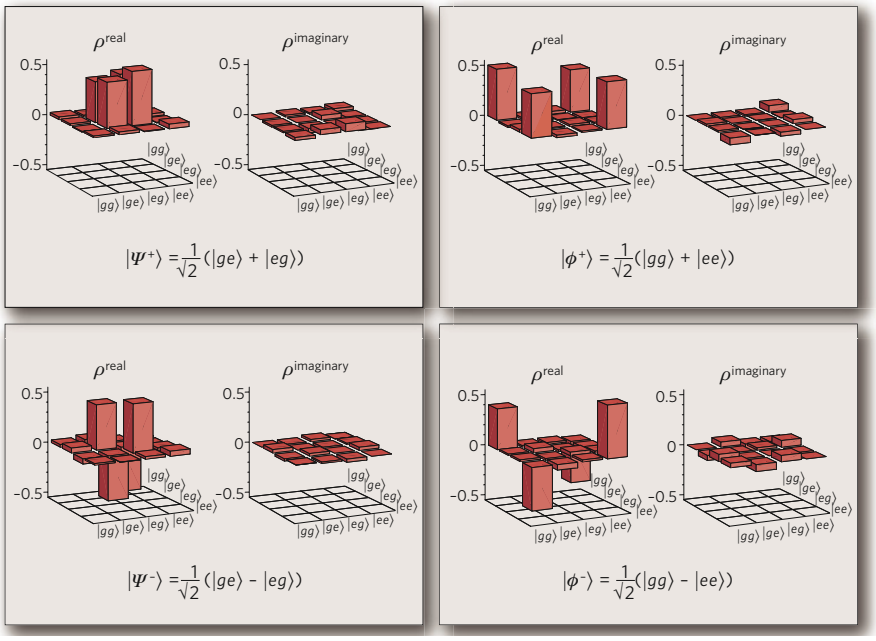


Figure 5 | Measured density matrices of Bell states. The real (left) and imaginary (right) parts of the density matrices obtained for the Bell states $|\Psi^+\rangle$ (upper left), $|\Psi^-\rangle$ (lower left), $|\phi^+\rangle$ (upper right) and $|\phi^-\rangle$ (lower right) prepared deterministically with two trapped $^{40}Ca^+$ ions are shown. The states were analysed by using quantum-state tomography, a technique that provides all of the necessary information to reconstruct the corresponding density matrix[8]. More specifically, the density matrix for a single qubit can be represented by $\rho = \frac{1}{2}(I + \Sigma_i \langle\sigma_i\rangle \sigma_i)$, where σ_i is a Pauli matrix, $i = x, y, z$ and I is the identity matrix. Measurements project a qubit onto its energy eigenstates, which is equivalent to measuring $\langle\sigma_z\rangle$. To determine $\langle\sigma_{x,y}\rangle$, an additional rotation of the Bloch sphere is applied before the measurement. The tomography procedure can be extended to N qubits, requiring of the order of 4^N expectation values to be measured. Owing to statistical errors, the experimentally measured expectation values can result in unphysical elements in the density matrix (with negative eigenvalues). This outcome is avoided by fitting the measured expectation values by using a maximum-likelihood method and then finding the most likely density matrix that describes the state[28].

Prospects

Although the basic elements of quantum computation have been demonstrated with atomic ions, operation errors must be significantly reduced and the number of ion qubits must be substantially increased if quantum computation is to be practical. Nevertheless, before fidelities and qubit numbers reach those required for a useful factoring machine, worthwhile quantum simulations might be realized.

More ion qubits and better fidelity

To create many-ion entangled states, there are two important goals: improving gate fidelity, and overcoming the additional problems that are associated with large numbers of ions. For fault-tolerant operation, a reasonable guideline is to assume that the probability of an error occurring during a single gate operation should be of the order of 10^{-4} or lower. An important benchmark is the fidelity of two-qubit gates. The best error probability achieved so far is approximately 10^{-2}, which was inferred from the fidelity of Bell-state generation[63]. In general, it seems that gate fidelities are compromised by limited control of classical components (such as fluctuations in the laser-beam intensity at the positions of the ions) and by quantum limitations (such as decoherence caused by spontaneous emission)[64]. These are daunting technical problems; however, eventually, with sufficient care and engineering expertise, these factors are likely to be suppressed.

The multiqubit operations discussed in this review rely on the ability to isolate spectrally a single mode of the motion of an ion. Because there are $3N$ modes of motion for N trapped ions, as N becomes large, the mode spectrum becomes so dense that the gate speeds must be significantly reduced to avoid off-resonance coupling to other modes. Several proposals have been put forward to circumvent this problem[65,66]. Alternatively, a way to solve this problem with gates that have been demonstrated involves distributing the ions in an array of multiple trap zones[18,67–69] (Fig. 6a). In this architecture, multiqubit gate operations could be carried out on a relatively small number of ions in multiple processing zones. Entanglement could be distributed between these zones by physically moving the ions[18,68,69] or by optical means[25,67,70–72]. For quantum communication over large distances, optical distribution seems to be the only practical choice; for experiments in which local entanglement is desirable, moving ions is also an option.

Examples of traps that could be used for scaling up the number of ions used in an algorithm are shown in Fig. 6b. Ions can be moved between zones by applying appropriate control electric potentials to the various electrode segments[46,73–75]. Individual ions have been moved ~1 mm in ~50 μs without loss of coherence; the excitation of the ion's motion (in its local well) was less than one quantum[73]. Multiple ions present in a single zone can be separated[46,73] by inserting an electric potential 'wedge' between the ions. In the teleportation experiment by the NIST group[46], two ions could be separated from a third in ~200 μs, with negligible excitation of the motional mode used for subsequent entangling operations between the two ions. This absence of motional excitation meant that an additional entangling-gate operation on the separated ions could be implemented with reasonable fidelity. For algorithms that operate over long time periods, the ions' motion will eventually become excited as a result of transportation and background noise from electric fields. To counteract this problem, additional laser-cooled ions could be used to cool the qubits 'sympathetically' (Fig. 6a). These 'refrigerator' ions could be identical to the qubit ions[76], of a different isotope[77] or of a different species[60,78]. They could also aid in detection and state preparation (described earlier).

For all multiqubit gates that have been implemented so far, the speeds are proportional to the frequencies of the modes of the ions, which scale as $1/d_{qe}^2$, where d_{qe} is the distance of the ion to the nearest electrode. Therefore, it would be valuable to make traps as small as possible. Many groups have endeavoured to achieve this, but they have all observed significant heating of the ions, compromising gate fidelity. The heating is anomalously large compared with that expected to result from thermal noise, which arises from resistance in, or coupled to, the trap electrodes[18,79–83]. It scales approximately as $1/d_{qe}^4$ (refs 18, 79–83), which is consistent with the presence of independently fluctuating potentials on electrode patches, the extent of which is small compared with d_{qe} (ref. 79). The source of the heating has yet to be understood, but recent experiments[80,82] indicate that it is thermally activated and can be significantly suppressed by operating at low temperature.

For large trap arrays, a robust means of fabrication will be required, as well as means of independently controlling a very large number of electrodes. Microelectromechanical systems (MEMS) fabrication technologies can be used for monolithic construction[83,84], and trap structures can be further simplified by placing all electrodes in a plane[84,85]. To mitigate the problem of controlling many electrodes, it might be possible to incorporate 'on-board' electronics close to individual trap zones[86]. Laser beams must also be applied in several locations simultaneously, because it will be essential to carry out parallel operations when implementing complex algorithms. The recycling of laser beams can be used[86,87], but the overall laser power requirements will still increase. If gates are implemented by using stimulated-Raman transitions, then a

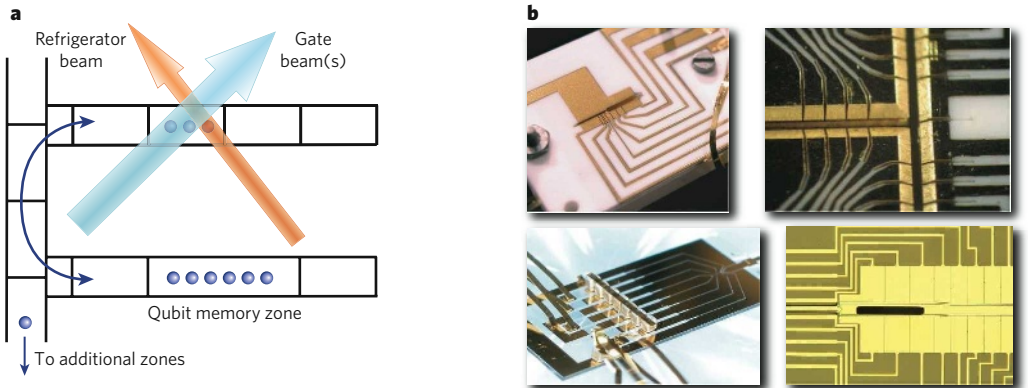


Figure 6 | Multizone trap arrays. a, A schematic representation of a multizone trap array is shown. Each control electrode is depicted as a rectangle. Ions (blue circles) can be separated and moved to specific zones, including a memory zone, by applying appropriate electrical potentials. Because the ions' motion will become excited as a result of transport (bidirectional arrow) and noisy ambient electric fields, refrigerator ions (red; which are cooled by the red laser beam) are used to cool the ions before gate operations, which are implemented with the blue laser beam. **b**, Examples of the electrode configurations of trap arrays are shown. In the upper left is a two-layer, six-zone linear trap in which entangled ions can be separated and used for algorithm demonstrations, including teleportation[46] (width of narrow slot (where the ions are located) = 200 μm). In the upper right is a three-layer, two-dimensional multizone trap that can be used to switch ion positions[99] (width of slot = 200 μm). In the lower left is a single-zone trap in which all of the electrodes lie in a single layer; this design considerably simplifies fabrication[85]. In the lower right is a single-layer, linear multizone trap fabricated on silicon (width of open slot for loading ions ≃ 95 μm), which can enable electronics to be fabricated on the same substrate that contains the trap electrodes. (Image courtesy of R. Slusher, Georgia Tech Research Institute, Atlanta).

high laser-beam intensity will also be needed to suppress spontaneous emission decoherence to fault-tolerant levels[64]. Detection will also need to be implemented simultaneously in several locations. This issue might be resolved by coupling on-board detectors or other forms of miniature integrated optics to optical fibres.

Future applications

In the early 1980s, Feynman suggested that one quantum system could perhaps be used to simulate another[5]. This simulation could be accomplished efficiently with a large-scale quantum computer. But before this goal is reached, it might be possible to take advantage of the fact that current logic gates are implemented by hamiltonians that can be used to simulate interactions in other systems. A simple example was mentioned earlier in the discussion of spectroscopy with cat states; these experiments simulate the action of electron, photon and atom Mach–Zehnder interferometers that incorporate entangling beam splitters[55]. A more interesting prospect is that the gate hamiltonians might be applied in a strategic way to simulate specific many-body hamiltonians. The basic idea can be considered by noting that the two-ion phase gate (Fig. 3) can be written in the form $R_{Z1}R_{Z2}e^{-i\xi\sigma_{z1}\sigma_{z2}}$, where R_{Z1} and R_{Z2} are rotations about the z axis and ξ is the strength of coupling. Therefore, up to an overall rotation on the qubits, the gate implements the hamiltonian $H=\hbar\kappa\sigma_{z1}\sigma_{z2}$, a spin–spin interaction between the two addressed spins, where κ is the strength of the interaction and $\hbar$ is $h/2\pi$ (and h is Planck's constant). By extending these couplings to many ion qubits in an ensemble, Ising-type spin hamiltonians could, for example, be implemented[88–91]. The interactions between ion pairs could be applied in a stepwise manner but might also be implemented simultaneously, thereby increasing efficiency. Although simulating specific many-body hamiltonians is a challenge given current experimental capabilities, even with a relatively small number of ions, interesting phenomena such as quantum phase transitions might be observable.

Conclusion

As researchers progress towards generating a large-scale quantum-information-processing device, it might be possible to shed light on more fundamental issues of decoherence and why many-particle states with the quantum attributes of Schrödinger's cat are not observed. If it is possible to continue scaling up such devices to a large size, the issue of the absence of cat states becomes more pressing. For example, suppose that, in the future, large-N-qubit cat states in the form of equation (3) can be made. Then, this cat state for N qubits can be rewritten as $|\Psi\rangle=\frac{1}{\sqrt{2}}(|g\rangle_j\pi_{k\neq j}^N|g\rangle_k+|e\rangle_j\pi_{k\neq j}^N|e\rangle_k)$, where the jth qubit has been (arbitrarily) singled out and k represents the other qubits. For large N, this wavefunction has the attributes of Schrödinger's cat in the sense that the states of a single two-level quantum system (the jth qubit) are correlated with states that have macroscopically distinct polarizations. If generating such states is successful, then the existence of, in essence, Schrödinger's cats will have been shown. Such states are, however, more sensitive to the effects of phase decoherence[56], but this seems to be a technical, not a fundamental, problem. Therefore, if it becomes impossible to make such states or to build a large-scale quantum computer for non-technical reasons, this failure might indicate some new physics. ■

1. Ramsey, N. F. *Molecular Beams* (Clarendon, London, 1956).
2. Freedman, S. F. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28,** 938–941 (1972).
3. Aspect, A., Grangier, P. & Roger, G. Experimental tests of realistic local theories via Bell's theorem. *Phys. Rev. Lett.* **47,** 460–463 (1981).
4. Bell, J. S. *Speakable and Unspeakable in Quantum Mechanics* (Cambridge Univ. Press, Cambridge, UK, 1987).
5. Feynman, R. P. Simulating physics with computers. *Int. J. Theoret. Phys.* **21,** 467–468 (1982).
6. Deutsch, D. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A* **400,** 97–117 (1985).
7. Shor, P. W. Algorithms for quantum computation: discrete logarithms and factoring. In *Proc. Annu. Symp. Found. Comput. Sci.* 124–134 (1994).
8. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, UK, 2000).
9. Cirac, J. I. & Zoller, P. Quantum computations with cold trapped ions. *Phys. Rev. Lett.* **74,** 4091–4094 (1995).
10. Monroe, C. Quantum information processing with atoms and photons. *Nature* **416,** 238–246 (2002).
11. Dehmelt, H. Experiments with an isolated subatomic particle at rest. *Rev. Mod. Phys.* **62,** 525–530 (1990).
12. Paul, W. Electromagnetic traps for charged and neutral particles. *Rev. Mod. Phys.* **62,** 531–540 (1990).
13. Bollinger, J. J., Heinzen, D. J., Itano, W. M., Gilbert, S. L. & Wineland, D. J. A 303-MHz frequency standard based on trapped $^9Be^+$ ions. *IEEE Trans. Instrum. Meas.* **40,** 126–128 (1991).
14. Fisk, P. T. H. *et al.* Very high q microwave spectroscopy on trapped $^{171}Yb^+$ ions: application as a frequencystandard. *IEEE Trans. Instrum. Meas.* **44,** 113–116 (1995).
15. Blatt, R., Häffner, H., Roos, C., Becher, C. & Schmidt-Kaler, F. in *Quantum Entanglement and Information Processing: Les Houches Session LXXIX* (eds Estève, D. Raimond, J.-M. & Dalibard, J.) 223–260 (Elsevier, Amsterdam, 2004).
16. Wineland, D. J. in *Quantum Entanglement and Information Processing: Les Houches Session LXXIX* (eds Estève, D. Raimond, J.-M. & Dalibard, J.) 261–293 (Elsevier, Amsterdam, 2004).
17. Leibfried, D., Blatt, R., Monroe, C. & Wineland, D. Quantum dynamics of single trapped ions. *Rev. Mod. Phys.* **75,** 281–324 (2003).
18. Wineland, D. J. *et al.* Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl Inst. Stand. Technol.* **103,** 259–328 (1998).
19. Diedrich, F., Bergquist, J. C., Itano, W. M. & Wineland, D. J. Laser cooling to the zero-point energy of motion. *Phys. Rev. Lett.* **62,** 403–406 (1989).
20. Dehmelt, H. G. Mono-ion oscillator as potential ultimate laser frequency standard. *IEEE Trans. Instrum. Meas.* **31,** 83–87 (1982).
21. Monroe, C., Meekhof, D. M., King, B. E., Itano, W. M. & Wineland, D. J. Demonstration of a fundamental quantum logic gate. *Phys. Rev. Lett.* **75,** 4714–4717 (1995).
22. Schmidt-Kaler, F. *et al.* Realization of the Cirac–Zoller controlled-NOT quantum gate. *Nature* **422,** 408–411 (2003).
23. Schmidt-Kaler, F. *et al.* How to realize a universal quantum gate with trapped ions. *Appl. Phys. B* **77,** 789–796 (2003).
24. Riebe, M. *et al.* Process tomography of ion trap quantum gates. *Phys. Rev. Lett.* **97,** 220407 (2006).
25. Moehring, D. L. *et al.* Entanglement of single-atom quantum bits at a distance. *Nature* **449,** 68–71 (2007).
26. Turchette, Q. A. *et al.* Deterministic entanglement of two trapped ions. *Phys. Rev. Lett.* **81,** 3631–3634 (1998).
27. Rowe, M. A. *et al.* Experimental violation of a Bell's inequality with efficient detection. *Nature* **409,** 791–794 (2001).
28. Roos, C. F. *et al.* Bell states of atoms with ultralong life times and their tomographic state analysis. *Phys. Rev. Lett.* **92,** 220402 (2004).
29. Sackett, C. A. *et al.* Experimental entanglement of four particles. *Nature* **404,** 256–259 (2000).
30. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23,** 880–884 (1969).
31. Moehring, D. L., Madsen, M. J., Blinov, B. B. & Monroe, C. Experimental Bell inequality violation with an atom and a photon. *Phys. Rev. Lett.* **93,** 090410 (2004).
32. Schrödinger, E. Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften* **23,** 807–812 (1935).
33. Greenberger, D. M., Horne, M. A. & Zeilinger, A. in *Going Beyond Bell's Theorem* (ed. Kafatos, M.) 69–72 (Kluwer Academic, Dordrecht, 1989).
34. DiVincenzo, D. P. & Shor, P. W. Fault-tolerant error correction withefficient quantum codes. *Phys. Rev. Lett.* **77,** 3260–3263 (1996).
35. Steane, A. M. Error correcting codes in quantum theory. *Phys. Rev. Lett.* **77,** 793–797 (1996).
36. Bollinger, J. J., Itano, W. M., Wineland, D. J. & Heinzen, D. J. Optimal frequency measurements with maximally correlated states. *Phys. Rev. A* **54,** R4649–R4652 (1996).
37. Leibfried, D. *et al.* Toward Heisenberg-limited spectroscopy with multiparticle entangled states. *Science* **304,** 1476–1478 (2004).
38. Leibfried, D. *et al.* Creation of a six-atom 'Schrödinger cat' state. *Nature* **438,** 639–642 (2005).
39. Roos, C. F. *et al.* Control and measurement of three-qubit entangled states. *Science* **304,** 1478–1480 (2004).
40. Dür, W., Vidal, G. & Cirac, J. I. Three qubits can be entangled in two inequivalent ways. *Phys. Rev. A* **62,** 062314 (2000).
41. Häffner, H. *et al.* Scalable multiparticle entanglement of trapped ions. *Nature* **438,** 643–646 (2005).
42. Deutsch, D. & Jozsa, R. Rapid solution of problems by quantum computation. *Proc. R. Soc. Lond. A* **439,** 553–558 (1992).
43. Chuang, I. L. *et al.* Experimental realization of a quantum algorithm. *Nature* **393,** 143–146 (1998).
44. Gulde, S. *et al.* Implementation of the Deutsch–Jozsa algorithm on an ion-trap quantum computer. *Nature* **421,** 48–50 (2003).
45. Bennett, C. H. *et al.* Teleporting an unknown quantum state via dual classical and Einstein–Podolsky-Rosen channels. *Phys. Rev. Lett.* **70,** 1895–1899 (1993).
46. Barrett, M. D. *et al.* Deterministic quantum teleportation of atomic qubits. *Nature* **429,** 737–739 (2004).
47. Riebe, M. *et al.* Deterministic quantum teleportation with atoms. *Nature* **429,** 734–737 (2004).
48. Reichle, R. *et al.* Experimental purification of two-atom entanglement. *Nature* **443,** 838–841 (2006).
49. Chiaverini, J. *et al.* Realization of quantum error correction. *Nature* **432,** 602–605 (2004).
50. Chiaverini, J. *et al.* Implementation of the semiclassical quantum Fourier transform in a scalable system. *Science* **308,** 997–1000 (2005).
51. Grover, L. K. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.* **79,** 325–328 (1997).
52. Wineland, D. J., Bollinger, J. J., Itano, W. M., Moore, F. L. & Heinzen, D. J. Spin squeezing and reduced quantum noise in spectroscopy. *Phys. Rev. A* **46,** R6797–R6800 (1992).

53. Itano, W. M. *et al.* Quantum projection noise: population fluctuations in two-level systems. *Phys. Rev. A* **47,** 3554–3570 (1993).
54. Meyer, V. *et al.* Experimental demonstration of entanglement-enhanced rotation angle estimation using trapped ions. *Phys. Rev. Lett.* **86,** 5870–5873 (2001).
55. Leibfried, D. *et al.* Trapped-ion quantum simulator: experimental application to nonlinear interferometers. *Phys. Rev. Lett.* **89,** 247901 (2002).
56. Huelga, S. F. *et al.* Improvement of frequency standards with quantum entanglement. *Phys. Rev. Lett.* **79,** 3865–3868 (1997).
57. André, A., Sørensen, A. S. & Lukin, M. D. Stability of atomic clocks based on entangled atoms. *Phys. Rev. Lett.* **92,** 230801 (2004).
58. Schaetz, T. *et al.* Enhanced quantum state detection efficiency through quantum information processing. *Phys. Rev. Lett.* **94,** 010501 (2005).
59. Hume, D. B., Rosenband, T. & Wineland, D. J. High-fidelity adaptive qubit detection through repetitive quantum nondemolition measurements. *Phys. Rev. Lett.* **99,** 120502 (2007).
60. Rosenband, T. *et al.* Frequency ratio of Al^+ and Hg^+ single-ion optical clocks; metrology at the 17th decimal place. *Science* **319,** 1808–1812 (2008).
61. Guerlin, C. *et al.* Progressive field-state collapse and quantum non-demolition photon counting. *Nature* **448,** 889–894 (2007).
62. Roos, C. F., Chwalla, M., Kim, K., Riebe, M. & Blatt, R. 'Designer atoms' for quantum metrology. *Nature* **443,** 316–319 (2006).
63. Benhelm, J., Kirchmair, G., Roos, C. F. & Blatt, R. Towards fault-tolerant quantum computing with trapped ions. *Nature Phys.* **4,** 463–466 (2008).
64. Ozeri, R. *et al.* Errors in trapped-ion quantumgates due to spontaneous photon scattering. *Phys. Rev. A* **75,** 042329 (2007).
65. Zhu, S.-L., Monroe, C. & Duan, L.-M. Arbitrary-speed quantum gates within large ion crystals through miminum control of laser beams. *Europhys. Lett.* **73,** 485–491 (2006).
66. Duan, L.-M. Scaling ion trap quantum computation through fast quantum gates. *Phys. Rev. Lett.* **93,** 100502 (2004).
67. DeVoe, R. G. Elliptical ion traps and trap arrays for quantum computation. *Phys. Rev. A* **58,** 910–914 (1998).
68. Cirac, J. I. & Zoller, P. A scalable quantum computer with ions in an array of microtraps. *Nature* **404,** 579–581 (2000).
69. Kielpinski, D., Monroe, C. & Wineland, D. J. Architecture for a large-scale ion-trap quantum computer. *Nature* **417,** 709–711 (2002).
70. Cirac, I., Zoller, P., Kimble, J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78,** 3221–3224 (1997).
71. Duan, L.-M. & Kimble, H. J. Scalable photonic quantum computation through cavity-assisted interactions. *Phys. Rev. Lett.* **92,** 127902 (2004).
72. Duan, L.-M. *et al.* Probabilistic quantum gates between remote atoms through interference of optical frequency qubits. *Phys. Rev. A* **73,** 062324 (2006).
73. Rowe, M. *et al.* Transport of quantum states and separation of ions in a dual rf ion trap. *Quantum Inform. Comput.* **2,** 257–271 (2002).
74. Hucul, D. *et al.* On the transport of atomic ions in linear and multidimensional trap arrays. Preprint at <http://arxiv.org/abs/quant-ph/0702175> (2007).
75. Huber, G. *et al.* Transport of ions in a segmented linear Paul trap in printed-circuit-board technology. *New J. Phys.* **10,** 013004 (2008).
76. Rohde, H. *et al.* Sympathetic ground-state cooling and coherent manipulation with two-ion crystals. *J. Opt. Soc. Am. B* **3,** S34–S41 (2001).
77. Blinov, B. B. *et al.* Sympathetic cooling of trapped Cd^+ isotopes. *Phys. Rev. A* **65,** 040304 (2002).
78. Barrett, M. D. *et al.* Sympathetic cooling of $^9Be^+$ and $^{24}Mg^+$ for quantum logic. *Phys. Rev. A* **68,** 042302 (2003).
79. Turchette, Q. A. *et al.* Heating of trapped ions from the quantum ground state. *Phys. Rev. A* **61,** 063418 (2000).
80. Deslauriers, L. *et al.* Scaling and suppression of anomalous heating in ion traps. *Phys. Rev. Lett.* **97,** 103007 (2006).
81. Leibrandt, D., Yurke, B. & Slusher, R. Modeling ion trap thermal noise decoherence. *Quant. Inform. Comput.* **7,** 52–72 (2007).
82. Labaziewicz, J. *et al.* Suppression of heating rates in cryogenic surface-electrode ion traps. *Phys. Rev. Lett.* **100,** 013001 (2008).
83. Stick, D. *et al.* Ion trap in a semiconductor chip. *Nature Phys.* **2,** 36–39 (2006).
84. Chiaverini, J. *et al.* Surface-electrode architecture for ion-trap quantum information processing. *Quantum Inform. Comput.* **5,** 419–439 (2005).
85. Seidelin, S. *et al.* Microfabricated surface-electrode ion trap for scalable quantum information processing. *Phys. Rev. Lett.* **96,** 253003 (2006).
86. Kim, J. *et al.* System design for large-scale ion trap quantum information processor. *Quant. Inform. Comput.* **5,** 515–537 (2005).
87. Leibfried, D., Knill, E., Ospelkaus, C. & Wineland, D. J. Transport quantum logic gates for trapped ions. *Phys. Rev. A* **76,** 032324 (2007).
88. Wunderlich, C. & Balzer, C. Quantum measurements and new concepts for experiments with trapped ions. *Adv. At. Mol. Opt. Phys.* **49,** 293–376 (2003).
89. Porras, D. & Cirac, J. I. Quantum manipulation of trapped ions in two dimensional Coulomb crystals. *Phys. Rev. Lett.* **96,** 250501 (2006).
90. Taylor, J. M. & Calarco, T. Wigner crystals of ions as quantum hard drives. Preprint at <http://arxiv.org/abs/0706.1951> (2007).
91. Chiaverini, J. & Lybarger Jr, W. E. Laserless trapped-ion quantum simulations without spontaneous scattering using microtrap arrays. *Phys. Rev. A* **77,** 022324 (2008).
92. Mølmer, K. & Sørensen, A. Multiparticle entanglement of hot trapped ions. *Phys. Rev. Lett.* **82,** 1835–1838 (1999).
93. Milburn, G. J., Schneider, S. & James, D. F. Ion trap quantum computing with warm ions. *Fortschr. Physik* **48,** 801–810 (2000).
94. Solano, E., de Matos Filho, R. L. & Zagury, N. Mesoscopic superpositions of vibronic collective states of *N* trapped ions. *Phys. Rev. Lett.* **87,** 060402 (2001).
95. Leibfried, D. *et al.* Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. *Nature* **422,** 412–415 (2003).
96. Haljan, P. C. *et al.* Entanglement of trapped-ion clock states. *Phys. Rev. A* **72,** 062316 (2005).
97. Home, J. P. *et al.* Deterministic entanglement and tomography of ion spin qubits. *New J. Phys.* **8,** 188 (2006).
98. Matsukevich, D. N., Maunz, P., Moehring, D. L., Olmschenk, S. & Monroe, C. Bell inequality violation with two remote atomic qubits. *Phys. Rev. Lett.* **100,** 150404 (2008).
99. Hensinger, W. K. *et al.* T-junction ion trap array for two-dimensional ion shuttling storage, and manipulation. *Appl. Phys. Lett.* **88,** 034101 (2006).

Acknowledgements We thank H. Häffner, J. Home, E. Knill, D. Leibfried, C. Roos and P. Schmidt for comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to R.B. (Rainer.Blatt@uibk.ac.at).

NATURE|Vol 453|19 June 2008|doi:10.1038/nature07126

Quantum coherence and entanglement with ultracold atoms in optical lattices

Immanuel Bloch[1]

At nanokelvin temperatures, ultracold quantum gases can be stored in optical lattices, which are arrays of microscopic trapping potentials formed by laser light. Such large arrays of atoms provide opportunities for investigating quantum coherence and generating large-scale entanglement, ultimately leading to quantum information processing in these artificial crystal structures. These arrays can also function as versatile model systems for the study of strongly interacting many-body systems on a lattice.

Recent advances in the laser cooling of neutral (uncharged) atoms and the creation of ultracold quantum gases[1] have opened up intriguing possibilities for the quantum manipulation of arrays of neutral atoms. Around 15–20 years ago, spectacular progress was made on the trapping and spectroscopy of single particles, and researchers concentrated on studying such single particles with ever-increasing precision. Now, researchers are building on these exquisite manipulation and trapping techniques to extend this control over larger arrays of particles. Not only can neutral atoms be trapped in microscopic potentials engineered by laser light[2–4], but the interactions between these particles can be controlled with increasing precision. Given this success, the creation of large-scale entanglement and the use of ultracold atoms as interfaces between different quantum technologies have come to the forefront of research, and ultracold atoms are among the 'hot' candidates for quantum information processing, quantum simulations and quantum communication.

Two complementary lines of research using ultracold atoms are dominating this field. In a bottom-up approach, arrays of atoms can be built up one by one. By contrast, a top-down approach uses the realization of degenerate ultracold bosonic[5–7] and fermionic[8–10] quantum gases as an alternative way of establishing large-scale arrays of ultracold atoms; this approach allows the creation of large numbers of neutral atoms, with almost perfect control over the motional and electronic degrees of freedom of millions of atoms with temperatures in the nanokelvin range. When such ultracold atoms are loaded into three-dimensional arrays of microscopic trapping potentials, known as optical lattices, the atoms are sorted in such a way that every lattice site is occupied by a single atom, for example, by strong repulsive interactions in the case of bosons or by Pauli blocking in the case of fermions. For bosons, this corresponds to a Mott insulating state[11–15], whereas for fermions a band insulating state is created[16], both of which form a highly regular, ordered, quantum register at close to zero kelvin. After initialization, the interactions and the states of the atoms are controlled to coax them into the correct — possibly entangled — macroscopic (many body) state to be used in quantum information processing, for example, or metrology at the quantum limit.

Ultracold atoms cannot yet rival the pristine control achieved using ion-trap experiments (see page 1008), but some key features nevertheless render them highly attractive. First, neutral atoms couple only weakly to the environment, allowing long storage and coherence times, even in the proximity of bulk materials; this feature has made them highly successful in the field of cavity quantum electrodynamics (see page 1023). Second, ultracold atoms in optical lattices form the only system so far in which a large number (up to millions) of particles can be initialized simultaneously. Eventually, any system proposed for quantum information processing will have to deal with such large arrays, and many of the perspectives (and difficulties) associated with these can already be tested using ultracold atoms today. Ultracold atoms have therefore also become promising candidates in a related line of research — quantum simulations[4,17–19] — in which highly controllable quantum matter is used to unravel some of the most intriguing questions in modern condensed-matter physics involving strongly correlated many-body quantum systems. In this review, I describe basic aspects of optical trapping and optical lattices. I then discuss novel state manipulation and entanglement schemes in optical lattices, and how these might be used to implement measurement-based quantum computing.

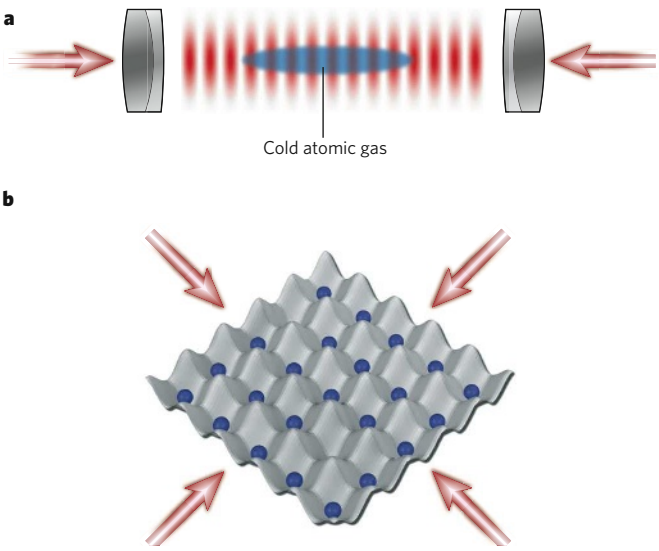


Figure 1 | Formation of optical lattices. a, An optical standing wave is generated by superimposing two laser beams. The antinodes (or nodes) of the standing wave act as a perfectly periodic array of microscopic laser traps for the atoms. The crystal of light in which the cold atoms can move and are stored is called an optical lattice. **b**, If several standing waves are overlapped, higher-dimensional lattice structures can be formed, such as the two-dimensional optical lattice shown here.

[1]Institut für Physik, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany.

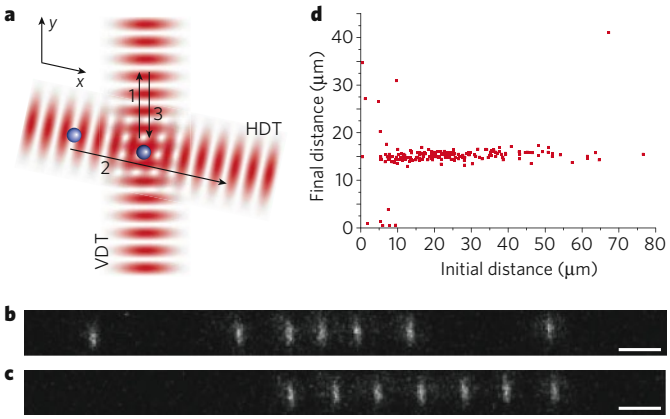


Figure 2 | Atom sorting in an optical lattice. a, Strings of atoms can be rearranged by using two crossed standing waves. Atoms can be moved independently in the horizontal or vertical direction by tuning the frequency difference of the counterpropagating laser beams, forming a single one-dimensional optical lattice. HDT, horizontal dipole trap; VDT, vertical dipole trap. **b**, Fluorescence image of the initial atom distribution on the lattice. Scale bar, 15 μm. **c**, Applying distance-control operations on six of the seven atoms creates a string of atoms with equidistant separation. This is carried out by moving the two standing waves through several sequences (for example, 1, 2, and then 3) as shown in **a**. The atoms follow the movement of the nodes of the lattices and can thereby be repositioned. Scale bar, 15 μm. **d**, For initial distances of the atoms larger than 10 μm, the atoms can be sorted to controlled separations of 15 μm. (Reproduced, with permission, from ref. 31.)

Optical trapping and optical lattices

Neutral atoms can be efficiently trapped by laser light thanks to the optical dipole force. This technique — in which cells can be manipulated with optical tweezers, without touching them — is widely used in biophysics. The basic principle is that a particle with an electric dipole moment **d** placed in an external electric field **E** experiences a potential energy: $V_{dip} = -\mathbf{d}\bullet\mathbf{E}$. In the case of an oscillating electric field, an oscillating electric dipole moment is induced, for example when laser light interacts with an atom. Such an induced dipole moment is proportional to the applied electric field strength and results in an optical potential that is generally proportional to the intensity of the applied light field. The optical potential can either be attractive or repulsive, depending on whether the frequency of the applied laser field is smaller or larger than the atomic resonance frequency[2].

Periodic potentials can be formed out of such optical potentials by interfering laser beams propagating along different directions. The resultant periodic pattern of bright and dark fringes is experienced by the atoms as a perfect array of potential maxima and minima in which they move. In the simplest case of two counterpropagating laser beams along the z axis, a periodic potential of the form $V_{lat} = V_0 \sin^2(2\pi z/\lambda)$ is created, with a periodicity of $\lambda/2$, where λ is the wavelength of the light field and V_0 is the potential depth of the lattice (Fig. 1a). By superimposing several of these standing-wave laser fields along different directions, it is possible to create lattice structures, in which atoms can be trapped, in one, two or three dimensions (Fig. 1b). For a three-dimensional lattice, each trapping site can be viewed as an almost perfect harmonic oscillator, with vibrational frequencies in the range of tens to hundreds of kilohertz. Such optical-lattice potentials offer huge flexibility in their design. For example, the potential depth can be changed along different directions independently, and the general lattice geometry can be controlled, for example by interfering laser beams at different angles. It has recently become possible to engineer spin-dependent lattice potentials, where different atomic spin states experience different periodic potentials[20,21], or superlattice structures composed of arrays of double wells[22–24]. When each of these double wells is filled with two atoms, they can mimic the behaviour of electronic double-quantum-dot systems[25–27], and similar strategies can be used to create protected and long-lived qubits and robust quantum gates. The additional strength of optical-lattice-based systems, however, lies in the fact that thousands of potential wells are present in parallel, each of which can be efficiently coupled with the neighbouring well to create massively parallel acting quantum gates.

Atom transport and state manipulation

One important challenge when dealing with ultracold atoms is keeping to a minimum any possible heating, because this could affect the motional or spin degrees of freedom. At the same time, atoms may need to be moved close together to initiate quantum gates between arbitrary pairs of atoms in the array. There has recently been an impressive advance in the control and movement of single atoms. A French research team has shown how a single atom, trapped in a dipole trap, can be moved in a two-dimensional plane in a highly controlled way with sub-micrometre spatial resolution[28]. The researchers also showed that atoms can be moved without detectable perturbation even if they are prepared in a coherent superposition of two internal spin states and when transferred from one dipole trap to another. In another approach, a team from the University of Bonn, Germany, used an 'atomic conveyer belt' to move and position atoms trapped in the nodes of a one-dimensional standing-wave light field[29]. By slightly tuning the frequency difference between the two counterpropagating laser fields, the standing wave can be turned into a 'walking wave', the motion of which the atoms closely track. By crossing two such conveyer belts along orthogonal directions, atoms can be actively sorted in an array. Such an 'atom sorting machine' has been used to sort a lattice randomly filled with seven atoms into a perfectly ordered string of equidistant single atoms[30,31] (Fig. 2). These impressive feats both contain crucial components for the controlled entanglement of atom pairs or strings of atoms in the lattice (discussed in the next section).

The control and imaging of single atoms in an optical lattice remains a huge challenge, but David Weiss and co-workers have recently shown how such imaging can work in a three-dimensional array of atoms[32]. By using a high-resolution optical lens, the researchers were able to image two-dimensional planes in a three-dimensional optical lattice

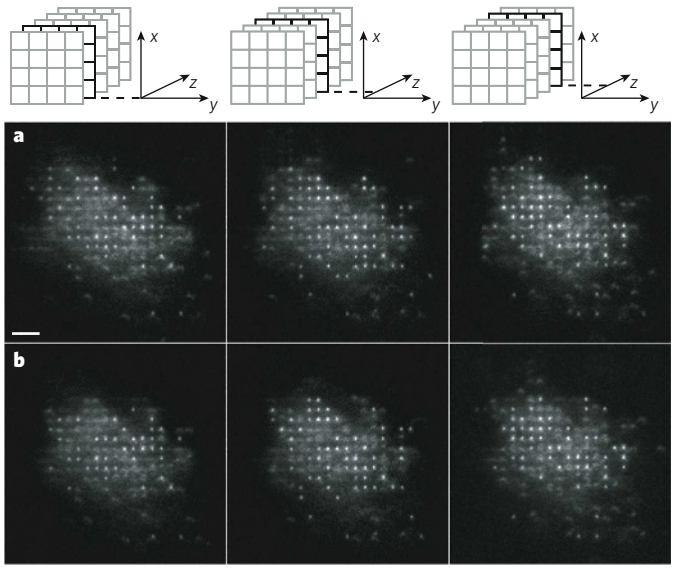


Figure 3 | Imaging of single atoms in a three-dimensional optical lattice. Up to 250 atoms are loaded from a magneto-optical trap into a three-dimensional optical lattice with a spacing of 4.9 μm. (Scale bar, 3 × 4.9 μm.) The atoms can be imaged by collecting their fluorescence light through a high-resolution objective lens. Different planes of the array can be targeted by focusing the imaging plane to different lattice planes (left to right). The same array of atoms can be imaged repeatedly while only minimally affecting the atom distribution in the lattice. Imaging was carried out along the z axis, at time $t = 0$ (**a**) and $t = 3$ s (**b**). (Reproduced, with permission, from ref. 32.)

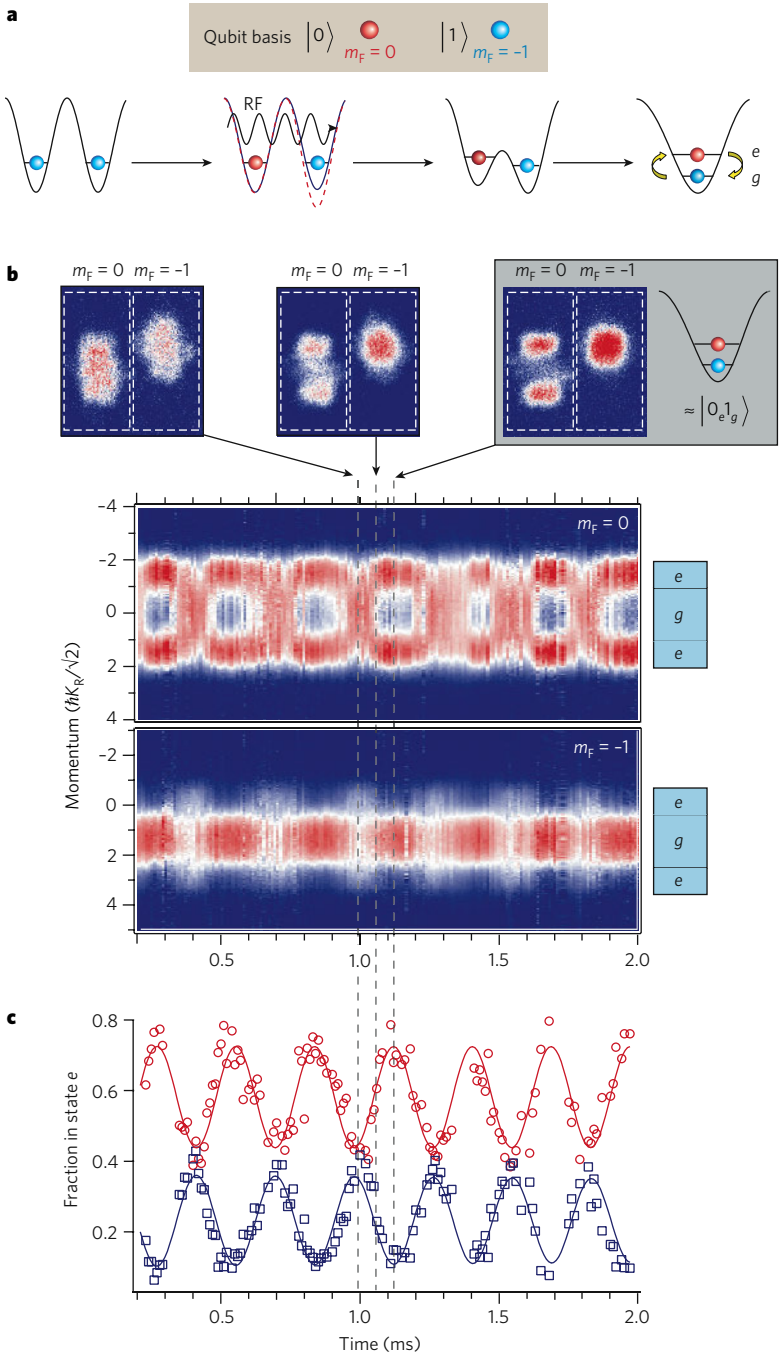


Figure 4 | Demonstration of a SWAP operation using exchange interactions. **a**, Using radio-frequency (RF) waves, two atoms in a double-well potential are brought into different spin states, denoted $|0\rangle$ (red) and $|1\rangle$ (blue), on different sides of the optical double-well potential. The two logical qubits, $|0\rangle$ and $|1\rangle$, are encoded in the electronic hyperfine states with angular momentum $m_F = 0$ and $m_F = -1$ of the atoms, respectively. When merging these into a single well, quantum-mechanical exchange interactions induce an oscillation between the spin populations in the lower and upper vibrational level. **b**, This oscillation can be revealed by using an adiabatic band mapping technique in which the population of different vibrational states is mapped onto different Brillouin zones after slowly turning off the lattice potential. (The Brillouin zones are given in units of $\hbar K_R\sqrt{2}$, where $\hbar K_R$ is the recoil momentum of the lattice photons. The colours reflect the number of atoms with this momentum, increasing from blue to red.) This technique allows the population of the vibrational states of a single lattice site to be measured in a spin-resolved way (upper images are examples of experimental results for the band mapping for three distinct times during the exchange oscillation cycle, with the times denoted by dashed lines through the lower images and images in **c**), revealing the exchange-induced spin dynamics (lower images, which show band mapping results taken at different times in the exchange oscillation cycle). The blue boxes indicate the momenta to which the different vibrational states, $|g\rangle$ and $|e\rangle$, are mapped. **c**, Multiple SWAP cycles are observed, by measuring the population in the excited vibrational state $|e\rangle$ over time (red, atoms in spin state $|0\rangle$; and blue, atoms in spin state $|1\rangle$). These show negligible decay during the oscillations, indicating the robust implementation of the two-qubit interaction. For half of a SWAP cycle, denoted as a $\sqrt{\text{SWAP}}$ operation, two atoms can be entangled to form a Bell pair. (Reproduced, with permission, from ref. 23.)

filled with up to 250 atoms loaded from a laser-cooled cloud of atoms (Fig. 3). To achieve such single-site and single-atom resolution, the team used a wider-spaced optical lattice with a periodicity of 4.9 μm, and the shallow depth of field of the optical detection allowed them to select a single lattice plane. Several groups are already trying to achieve such single-site and single-atom resolution[33–35] for tightly spaced lattices formed by counterpropagating laser beams in the optical regime, with a site spacing of only a few hundred nanometres. When such arrays are loaded from a degenerate bosonic or fermionic quantum gas, the lattice would be filled with hundreds of thousands of atoms, with each plane containing an array of typically 10,000 atoms that could be imaged and manipulated simultaneously.

Entangling neutral atoms

Storing, sorting and controlling atoms in a large-scale array of particles is only one part of the challenge; the other consists of entangling the particles to implement quantum gates or to generate multiparticle entangled resource states for quantum information processing. This requires precise control over the internal-state-dependent interactions between the particles in a lattice. Ideally, the interactions between any pair of atoms in the lattice should be controllable such that they could be coaxed into any desired quantum-mechanical superposition state. One approach is to use a single-atom read-and-write head, moving atoms in optical tweezers to the desired location to interact with other atoms. However, the transport takes precious time, during which harmful decoherence processes could destroy the fragile quantum coherence stored in the register.

Another possibility might be better adapted to the lattice system and takes advantage of the massive parallelism with which operations can be carried out. The interactions between neutral atoms are typically very short-ranged — they are known as 'contact interactions' — and only occur when two particles are brought together at a single lattice site, where they can directly interact. But when each atom is brought into contact with each of its neighbours, the collisions between the particles

can create a highly entangled multiparticle state[20,21], known as a 'cluster state'[36], which can be used as a resource state for quantum information processing. The superposition principle of quantum mechanics allows this to be achieved in a highly parallel way, using a state-dependent optical lattice, in which different atomic spin states experience different periodic potentials[20,21]. Starting from a lattice where each site is filled with a single atom, the atoms are first brought into a superposition of two internal spin states. The spin-dependent lattice is then moved in such a way that an atom in two different spin states splits up and moves to the left and right simultaneously so that it collides with its two neighbours. In a single operation, a whole string of atoms can thereby be entangled. However, if the initial string of atoms contained defects, an atom moving to the side may have no partner to collide with, so the length of the entangled cluster would be limited to the average length between two defects. The sorted arrays of atoms produced by an 'atomic sorting machine' could prove to be an ideal starting point for such collisional quantum gates, as the initial arrays are defect free. In addition, defects could be efficiently removed by further active cooling of the quantum gases in the lattice. Indeed, such cooling is necessary to enhance the regularity of the filling achieved with the current large-scale ensembles. Several concepts related to 'dark state' cooling methods from quantum optics and laser cooling could help in this case. The atoms could be actively cooled into the desired many-body quantum state, which is tailored to be non-interacting (that is, dark) with the applied cooling laser field[37,38].

When constructing such entangled states, the particles' many degrees of freedom can couple to the environment, leading to decoherence, which will destroy the complex quantum superpositions of the atoms. To avoid such decoherence processes, which affect the system more the larger it becomes, it is desirable to construct many-particle states, which are highly insensitive to external perturbations. Unfortunately, when using the outlined controlled-collisions scheme to create an atomic cluster state, the atomic qubits must be encoded in states that undergo maximal decoherence with respect to magnetic field fluctuations. Two recent experiments have shown how decoherence could be avoided, by implementing controlled exchange interactions between atoms[23,39]; this could lead to new ways of creating robust entangled states (discussed in the next section). Another way to avoid the problem of decoherence is to apply faster quantum gates, so more gate operations could be carried out within a fixed decoherence time. For the atoms of ultracold gases in optical lattices, Feshbach resonances[40,41] can be used to increase the collisional interactions and thereby speed up gate operations. However, the 'unitarity limit' in scattering theory does not allow the collisional interaction energy to be increased beyond the on-site vibrational oscillation frequency, so the lower timescale for a gate operation is typically a few tens of microseconds. Much larger interaction energies, and hence faster gate times, could be achieved by using the electric dipole–dipole interactions between polar molecules[42], for example, or Rydberg atoms[43,44]; in the latter case, gate times well below the microsecond range are possible. For Rydberg atoms, a phase gate between two atoms could be implemented by a dipole-blockade mechanism, which inhibits the simultaneous excitation of two atoms and thereby induces a phase shift in the two-particle state only when both atoms are initially placed in the same quantum state. The first signs of such a Rydberg dipole-blockade mechanism have been observed in mesoscopic cold and ultracold atom clouds[45–48], but it remains to be seen how well they can be used to implement quantum gates between two individual atoms. Rydberg atoms offer an important advantage for the entanglement of neutral atoms: they can interact over longer distances, and addressing single atoms in the lattice to turn the interactions between these two atoms on and off avoids the need for the atoms to move. In addition, the lattice does not have to be perfectly filled for two atoms to be entangled if their initial position is known before applying the Rydberg interaction.

Novel quantum gates via exchange interactions

Entangling neutral atoms requires state-dependent interactions. A natural way to achieve this is to tune the collisional interactions between atoms to different strengths for different spin states, or to allow explicitly only specific spin states into contact for controlled collisions. Another possibility is to exploit the symmetry of the underlying two-particle wavefunctions to create the desired gate operations, even in the case of completely spin-independent interactions between atoms. This principle lies at the heart of two experiments to control the spin–spin interactions between two particles using exchange symmetry[23,39,49], and builds on original ideas and experiments involving double quantum-dot systems[25,26].

Research teams at the National Institute of Standards and Technology (NIST) at Gaithersburg, Maryland, and the University of Mainz, Germany, have demonstrated such interactions for two atoms in a double-well potential. How do these exchange interactions arise, and how can they be used to develop primitives (or building blocks) for quantum information processing? As one of the fundamental principles of quantum mechanics, the total quantum state of two particles (used in two experiments) has to be either symmetrical in the case of bosons or antisymmetrical for fermions, with respect to exchange of the two particles. When trapped on a single lattice site, a two-particle bosonic wavefunction can be factored into a spatial component, which describes the positions of the two particles, and a spin component, which describes

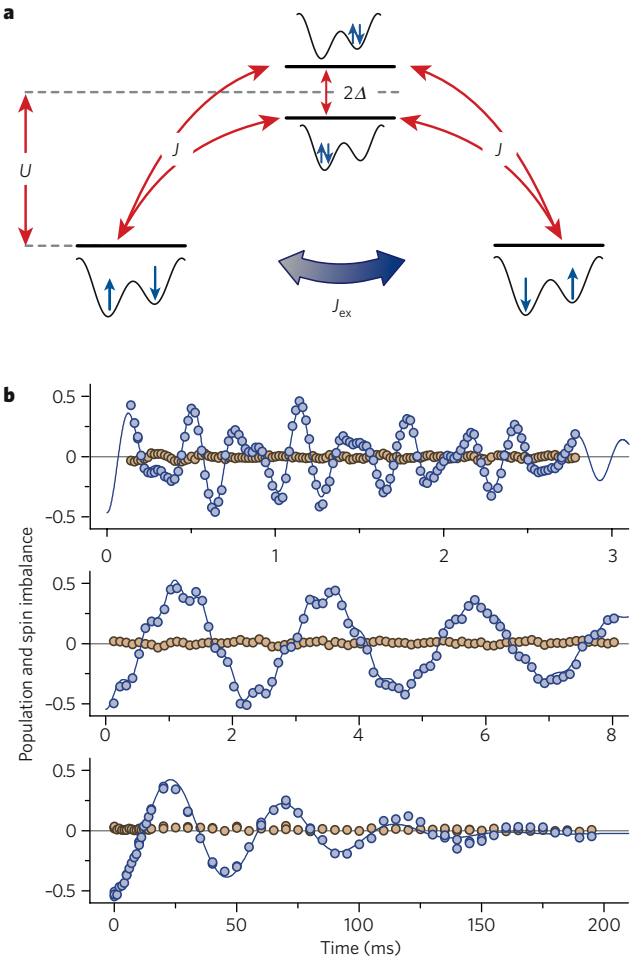


Figure 5 | Superexchange coupling between atoms on neighbouring lattice sites. a, Virtual hopping processes (left to right, and right to left) mediate an effective spin–spin interaction with strength J_{ex} between the atoms, which can be controlled in both magnitude and sign by using a potential bias Δ between the wells. U is the on-site interaction energy between the atoms on a single lattice site, and J is the single-particle tunnel coupling. **b**, The effective spin–spin interaction emerges when increasing the interaction U between the particles relative to their kinetic energy J (top to bottom). It can be observed in the time evolution of the magnetization dynamics in the double well. Blue circles indicate spin imbalance, and brown circles indicate population imbalance. The curves denote a fit to a theoretical model taking into account the full dynamics observed within the Hubbard model. (Reproduced, with permission, from ref. 39.)

their spin orientations. If the spatial wavefunction part is symmetrical with respect to particle exchange, the spin part must be symmetrical too, or they must both be antisymmetrical, so the total wavefunction always retains the correct symmetry. The two combinations, however, have different interaction energies: in the case of a symmetrical spatial wavefunction, both particles are more likely to be located in the same position, whereas for an antisymmetrical one they are never found at the same location. The former leads to strong collisional interactions between the particles, whereas the latter leads to a vanishing interaction energy. It is this energy difference between the 'singlet' (antisymmetrical) and 'triplet' (symmetrical) spin states that gives rise to an effective spin–spin interaction between the two particles.

When the NIST team placed two atoms onto a lattice site, with the spin-up particle in the vibrational ground state $|\uparrow, g\rangle$ and the spin-down particle in the first excited vibrational state $|\downarrow, e\rangle$, the effective spin interaction led to exchange oscillations between the qubit states $|\uparrow, g\rangle|\downarrow, e\rangle \Leftrightarrow |\downarrow, g\rangle|\uparrow, e\rangle$. In computer terminology this is called a SWAP operation and is one of the fundamental primitives of quantum computing[25]. In fact, the exchange operation allows for any transformations by an angle θ of the form $|a, b\rangle = \cos(\theta)|a, b\rangle + i\sin(\theta)|b, a\rangle$, for any spin state $|a\rangle$, $|b\rangle$ of the particles. When the SWAP operation is carried only halfway through, denoted by $\sqrt{\text{SWAP}}$, the two particles end up as an entangled Bell pair. The NIST researchers observed such SWAP operations by first preparing a $|\uparrow\rangle_L|\downarrow\rangle_R$ state configuration in the double-well potential (where L is the left well and R is the right well) and then actively deforming the double well, so both particles ended up on the same lattice site. Exchange oscillations then flipped the spin configurations over time; these were observed in the experiment over up to 12 SWAP cycles without any noticeable damping of the exchange oscillation signal[23] (Fig. 4). In the NIST experiments, the atoms had to be brought onto the same lattice site to initiate exchange interactions, but virtual tunnelling processes[24] can achieve this without moving the particles. In these processes, atoms constantly probe their neighbouring lattice site, after which either they or their neighbouring particle return to the original lattice site. Such a process can either leave the initial position of the atoms intact or swap them over, thereby giving rise to an effective spin–spin interaction between the two particles of the form $H_{\text{eff}} = -J_{\text{ex}}\mathbf{S}_i \bullet \mathbf{S}_j$, where $\mathbf{S}_i$ and $\mathbf{S}_j$ are the spin operators on neighbouring lattice sites i and j. Such 'super-exchange' interactions therefore do not require any direct wavefunction overlap of the two particles, as this overlap is established during the atoms' virtual hopping process. The strength and the sign of the coupling constant J_{ex} can be evaluated through second-order perturbation theory, resulting in $J_{\text{ex}} = 4J^2/U$, where J is the single-particle tunnelling coupling and U is the spin-independent interaction energy between two particles occupying the same lattice site[50–52]. The Mainz researchers could directly observe and control such superexchange spin couplings between two neighbouring atoms in the double-well potential created by an optical superlattice (Fig. 5). These controllable superexchange interactions form the basic building block of quantum magnetism in strongly correlated electronic media and give rise, for example, to the antiferromagnetic ordering of a two-component Fermi gas on a lattice[50]. For quantum information processing, they too can be used to implement SWAP operations, but their control over the spin states between pairs of atoms could find other uses as well. For example, by first creating an array of Bell pairs in optical superlattices using exchange interactions or spin-changing collisions[53], these Bell pairs could be connected to each other using Ising-type superexchange interactions to directly create cluster states or other useful resource states[54] (Fig. 6). Compared with the controlled-collision approaches, however, these cluster states can be encoded in substates with vanishing total magnetization and so could be more robust to global field fluctuations leading to decoherence.

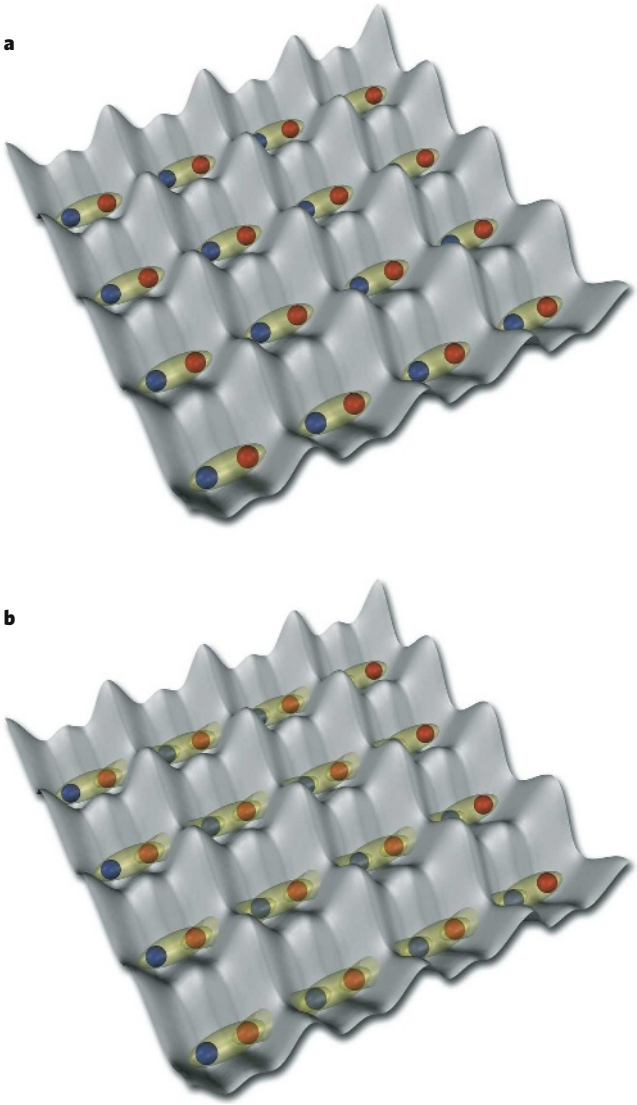


Figure 6 | Array of entangled Bell pairs obtained using optical superlattices. a, Using exchange-mediated $\sqrt{\text{SWAP}}$ operations, arrays of Bell pairs (yellow) consisting of two atoms in different spin states (red and blue) can be created in a massively parallel way. **b**, These two-particle entangled states can be extended to larger multiparticle entangled states, by using spin–spin interactions to connect atoms on the edges of a Bell pair (marked by additional yellow bonds between the edges of previously unconnected Bell pairs). Applying this operation additionally along the orthogonal direction leads to the creation of large two-dimensional cluster states or other useful entangled resource states[54].

Measurement-based quantum computing

In the field of quantum computing, there are several computational models, such as the quantum circuit model[55–57], adiabatic quantum computation[58], the quantum Turing machine[59,60], teleportation-based models[61–63] and the one-way quantum computer[64,65], giving rise to a large number of possibilities for how to carry out a quantum computation. In the circuit model, for example, information is processed through a series of unitary gate operations, after which the desired calculation result is obtained at the output. In the measurement-based one-way quantum computer, information is processed through a sequence of adaptive measurements on an initially prepared, highly entangled resource state. Measurement-based quantum computing (MBQC) lays out a wholly new concept for the practical implementation of quantum information processing that is extremely well suited to large arrays of particles, such as neutral atoms in optical lattices. First, a large, multiparticle, entangled resource state, such as a cluster state, is created by means of controlled collisions or the methods outlined above. A computational algorithm is then implemented by carrying out a sequence of adaptive single-particle measurements, together with local single-particle unitary operations (Fig. 7). The size of the initial entangled cluster is thereby crucial, as it determines the length of the calculation that can be carried out. Single-site addressing techniques that are currently being implemented in labs

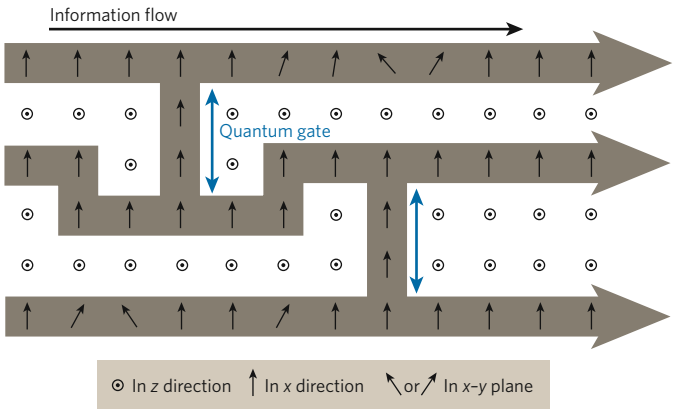


Figure 7 | Information processing in a one-way quantum computer. After initially creating a multiparticle entangled cluster state, a sequence of adaptive single-particle measurements is carried out. In each step of the computation, the measurement basis for the next qubit depends on the specific program and on the outcome of previous measurement results. Finally, after all the measurements have been carried out, the state of the system is given by $|\xi^{\{\alpha\}}\rangle|\Psi_{\rm out}^{\{\alpha\}}\rangle$, where the measured qubits are given by the product state $|\xi^{\{\alpha\}}\rangle$ and the final output state is $|\Psi_{\rm out}^{\{\alpha\}}\rangle$, which contains the computation result up to a unitary operation that depends on all of the previous measurement results, $\{\alpha\}$. The short black arrows in the figure denote the direction of the measurement basis for the corresponding qubit, and the large brown arrows indicate the directions of information flow. When measuring the qubits between two chains (blue arrows), a quantum gate is realized. (Reproduced, with permission, from ref. 36.)

could one day lead to cluster-state computing in lattice-based systems. Proof-of-principle demonstrations have already been carried out using photon-based cluster states[66,67], and the model could be implemented in any system consisting of an array of qubits.

So far, MBQC has already become a major research field, currently mainly driven by theory, with interdisciplinary connections to entanglement theory, graph theory, computational complexity, logic and statistical physics. Several fundamental questions regarding MBQC have now been answered, such as, which multiparticle entangled states can serve as 'universal resources'[68–71]. Universality in this context is defined as the ability to generate every possible quantum state from the resource through single-qubit operations alone. Using this definition, it can be shown that the two-dimensional cluster state is a universal resource state, whereas the one-dimensional cluster state is not. Furthermore, a universal resource state must be maximally entangled with respect to all types of entanglement measure. If this were not the case, there could be a state with a higher degree of entanglement that could not be generated from the resource state through single-qubit operations. Because single-qubit operations cannot add entanglement to the system, the initial state could not have been a universal resource state.

For MBQC to be implemented in practice, it is important to know how defects, such as missing atoms or doubly occupied sites, can limit its computational power. Active cooling of the lattice gases could help to reduce such defects[37,38], although a finite residual number of defects will always be present. Astonishingly, the computational power degrades sharply only when the number of defects is increased above the percolation threshold[72] of statistical physics. In addition, a cluster state can be a universal resource even in the presence of defects, although the location of the defects would need to be known in order to adapt a measurement sequence to them. In an effort to understand the computational power of MBQC, several teams have also shown how MBQC can be connected to other measurement-based quantum computing schemes, such as teleportation based ones[73–76].

Any real-world quantum computer will also need to overcome the adverse effects of decoherence arising from interactions with the environment, which affect the fragile quantum superpositions and the entangled many-body states in the system and result in errors in quantum computation. In the drive to create a scalable quantum computer, quantum error correction has a crucial role in correcting such errors[77], while maintaining the greater computational speed of a quantum computer over a classical computer. Quantum error correction allows an arbitrarily long quantum computation to be carried out with arbitrary accuracy, if the error level of the underlying operations is below a threshold value[78–80]. By combining topological error-correction schemes originating from Alexei Kitaev's toric code[81] and 'magic-state distillation' into the one-way quantum computer, it has recently been shown that an error threshold of up to 7.5×10^{-3} can be realized[82]. For a local model in two dimensions, in which only nearest-neighbour interactions between the particles are allowed, this is the highest threshold known, but it is still beyond the reach of current experiments.

Quantum simulations

Ultracold quantum gases in optical lattices are also being used to simulate the behaviour of strongly interacting electronic systems[4,17,19], where they might be able to shed light on complex problems emerging from condensed-matter physics. A prominent example is the Hubbard model, which forms a simple theoretical description of interacting fermions on a lattice. Although the basic hamiltonian for such a system can be easily written down, solving it is one of the hardest problems in condensed-matter physics. One problem that ultracold atoms might help to answer is whether a high-temperature superconducting phase can emerge from within the Hubbard model[83]. Such a scenario is widely thought to lie at the heart of the mystery of high-temperature superconductors[84]. A starting point for such studies could be an antiferromagnetically ordered gas of fermions, which after doping has been proposed to transform into a spin-liquid phase[84,85] that can support the formation of a high-temperature superconductor. Several research groups are currently trying to establish an antiferromagnetically ordered Mott insulator in fermionic atom clouds with two spin components. The temperature requirements to achieve this seem to be demanding[86], however, and progress will again depend on finding ways to cool the quantum gases within the lattice[37].

Outlook

From both an experimental and a theoretical point of view, optical lattices offer outstanding possibilities for implementing new designs for quantum information processing and quantum simulations. Some of the major experimental challenges in the field are lowering the temperatures of the lattice-based quantum gases and achieving single-site addressing, the latter being, for example, a crucial requirement for the MBQC model. Although there might be special situations in which this can be avoided, such addressability would provide a fresh impetus for the field of quantum simulations. Imagine being able to observe and control a spin system in two dimensions with 10,000 particles simultaneously in view, all with single-site and single-atom resolution. Observing dynamic evolutions in these systems, probing their spatial correlations and finally implementing quantum information processing in a truly large-scale system would all become possible. ■

1. Pitaevskii, L. & Stringari, S. *Bose–Einstein Condensation* (Oxford Univ. Press, Oxford, 2003).
2. Grimm, R., Weidemüller, M. & Ovchinnikov, Y. B. Optical dipole traps for neutral atoms. *Adv. At. Mol. Opt. Phys.* **42,** 95–170 (2000).
3. Jessen, P. S. & Deutsch, I. H. Optical lattices. *Adv. At. Mol. Opt. Phys.* **37,** 95–139 (1996).
4. Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold gases. Preprint at ⟨http://arxiv.org/abs/0704.3011⟩ (2007).
5. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. Observation of Bose–Einstein condensation in a dilute atomic vapor. *Science* **269,** 198–201 (1995).
6. Davis, K. B. *et al.* Bose–Einstein condensation in a gas of sodium atoms. *Phys. Rev. Lett.* **75,** 3969–3973 (1995).
7. Bradley, C. C., Sackett, C. A., Tollett, J. J. & Hulet, R. G. Evidence of Bose–Einstein condensation in an atomic gas with attractive interactions. *Phys. Rev. Lett.* **75,** 1687–1690 (1995).
8. DeMarco, B. & Jin, D. D. Onset of Fermi degeneracy in a trapped atomic gas. *Science* **285,** 1703–1706 (1999).
9. Schreck, F. *et al.* Quasipure Bose–Einstein condensate immersed in a Fermi Sea. *Phys. Rev. Lett.* **87,** 080403 (2001).
10. Truscott, A. G., Strecker, K. E., McAlexander, W. I., Partridge, G. P. & Hulet, R.G. Observation of Fermi pressure in a gas of trapped atoms. *Science* **291,** 2570–2572 (2001).

11. Fisher, M. P. A., Weichman, P. B., Grinstein, G. & Fisher, D. S. Boson localization and the superfluid-insulator transition. *Phys. Rev. B* **40,** 546–570 (1989).
12. Jaksch, D., Bruder, C., Cirac, J. I., Gardiner, C. W. & Zoller, P. Cold bosonic atoms in optical lattices. *Phys. Rev. Lett.* **81,** 3108–3111 (1998).
13. Greiner, M., Mandel, M. O., Esslinger, T., Hänsch, T. & Bloch, I. Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415,** 39–44 (2002).
14. Stöferle, T., Moritz, H., Schori, C., Köhl, M. & Esslinger, T. Transition from a strongly interacting 1D superfluid to a Mott insulator. *Phys. Rev. Lett.* **92,** 130403 (2004).
15. Spielman, I. B., Phillips, W. D. & Porto, J. V. The Mott insulator transition in two dimensions. *Phys. Rev. Lett.* **98,** 080404 (2007).
16. Köhl, M., Moritz, H., Stöferle, T., Günter, K. & Esslinger, T. Fermionic atoms in a three dimensional optical lattice: observing Fermi surfaces, dynamics, and interactions. *Phys. Rev. Lett.* **94,** 080403 (2005).
17. Jaksch, D. & Zoller, P. The cold atoms Hubbard toolbox. *Ann. Phys. (NY)* **315,** 52–79 (2005).
18. Bloch, I. Ultracold quantum gases in optical lattices. *Nature Phys.* **1,** 23–30 (2005).
19. Lewenstein, M. *et al.* Ultracold atomic gases in optical lattices: mimicking condensed matter physics and beyond. *Adv. Phys.* **56,** 243–379 (2007).
20. Jaksch, D., Briegel, H. J., Cirac, J. I., Gardiner, C. W. & Zoller, P. Entanglement of atoms via cold controlled collisions. *Phys. Rev. Lett.* **82,** 1975–1978 (1999).
21. Mandel, O. *et al.* Controlled collisions for multiparticle entanglement of optically trapped atoms. *Nature* **425,** 937–940 (2003).
22. Sebby-Strabley, J., Anderlini, M., Jessen, P. S. & Porto, J. V. Lattice of double wells for manipulating pairs of cold atoms. *Phys. Rev. A* **73,** 033605 (2006).
23. Anderlini, M. *et al.* Controlled exchange interaction between pairs of neutral atoms in an optical lattice. *Nature* **448,** 452–456 (2007).
24. Fölling, S. *et al.* Direct observation of second-order atom tunnelling. *Nature* **448,** 1029–1032 (2007).
25. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57,** 120–126 (1998).
26. Petta, J. R. *et al.* Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309,** 2180–2184 (2005).
27. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79,** 1217–1265 (2007).
28. Beugnon, J. *et al.* Two-dimensional transport and transfer of a single atomic qubit in optical tweezers. *Nature Phys.* **3,** 696–699 (2007).
29. Schrader, D. *et al.* A neutral atom quantum register. *Phys. Rev. Lett.* **93,** 150501 (2004).
30. Miroshnychenko, Y. *et al.* Precision preparation of strings of trapped neutral atoms. *New J. Phys.* **8,** 191 (2006).
31. Miroshnychenko, Y. *et al.* An atom-sorting machine. *Nature* **442,** 151 (2006).
32. Nelson, K. D., Li, X. & Weiss, D. S. Imaging single atoms in a three-dimensional array. *Nature Phys.* **3,** 556–560 (2007).
33. Cho, J. Addressing individual atoms in optical lattices with standing-wave driving fields. *Phys. Rev. Lett.* **99,** 020502 (2007).
34. Joo, J., Lim, Y. L., Beige, A. & Knight, P. L. Single-qubit rotations in two-dimensional optical lattices with multiqubit addressing. *Phys. Rev. A* **74,** 042344 (2006).
35. Gorshkov, A. V., Jiang, L., Greiner, M., Zoller, P. & Lukin, M. D. Coherent quantum optical control with subwavelength resolution. Preprint at ⟨http://arxiv.org/abs/0706.3879⟩ (2007).
36. Briegel, H. J. & Raussendorf, R. Persistent entanglement in arrays of interacting particles. *Phys. Rev. Lett.* **86,** 910–913 (2001).
37. Griessner, A., Daley, A. J., Clark, S. R., Jaksch, D. & Zoller, P. Dark-state cooling of atoms by superfluid immersion. *Phys. Rev. Lett.* **97,** 220403 (2006).
38. Griessner, A., Daley, A. J., Clark, S. R., Jaksch, D. & Zoller, P. Dissipative dynamics of atomic Hubbard models coupled to a phonon bath: dark state cooling of atoms within a Bloch band of an optical lattice. *New J. Phys.* **9,** 44 (2007).
39. Trotzky, S. *et al.* Time-resolved observation and control of superexchange interactions with ultracold atoms in optical lattices. *Science* **319,** 295–299 (2008).
40. Inouye, S. *et al.* Observation of Feshbach resonances in a Bose–Einstein condensate. *Nature* **392,** 151–154 (1998).
41. Courteille, P., Freeland, R. S., Heinzen, D. J., van Abeelen, F. A. & Verhaar, B. J. Observation of a Feshbach resonance in cold atom scattering. *Phys. Rev. Lett.* **81,** 69–72 (1998).
42. Micheli, A., Brennen, G. K. & Zoller, P. A toolbox for lattice-spin models with polar molecules. *Nature Phys.* **2,** 341–347 (2006).
43. Jaksch, D. *et al.* Fast quantum gates for neutral atoms. *Phys. Rev. Lett.* **85,** 2208–2211 (2000).
44. Lukin, M. D. *et al.* Dipole blockade and quantum information processing in mesoscopic atomic ensembles. *Phys. Rev. Lett.* **87,** 037901 (2001).
45. Tong, D. *et al.* Local blockade of Rydberg excitation in an ultracold gas. *Phys. Rev. Lett.* **93,** 063001 (2004).
46. Singer, K., Reetz-Lamour, M., Amthor, T., Marcassa, L. G. & Weidemüller, M. Suppression of excitation and spectral broadening induced by interactions in a cold gas of Rydberg atoms. *Phys. Rev. Lett.* **93,** 163001 (2004).
47. Liebisch, T. C., Reinhard, A., Berman, P. R. & Raithel, G. Atom counting statistics in ensembles of interacting Rydberg atoms. *Phys. Rev. Lett.* **95,** 253002 (2005).
48. Heidemann, R. *et al.* Evidence for coherent collective Rydberg excitation in the strong blockade regime. *Phys. Rev. Lett.* **99,** 163601 (2007).
49. Hayes, D., Julienne, P. S. & Deutsch, I. H. Quantum logic via the exchange blockade in ultracold collisions. *Phys. Rev. Lett.* **98,** 070501 (2007).
50. Auerbach, A. *Interacting Electrons and Quantum Magnetism* (Springer, New York, 2006).
51. Duan, L.-M., Demler, E. & Lukin, M. D. Controlling spin exchange interactions of ultracold atoms in an optical lattice. *Phys. Rev. Lett.* **91,** 090402 (2003).
52. Kuklov, A. B. & Svistunov, B. V. Counterflow superfluidity of two-species ultracold atoms in a commensurate optical lattice. *Phys. Rev. Lett.* **90,** 100401 (2003).
53. Widera, A. *et al.* Coherent collisional spin dynamics in optical lattices. *Phys. Rev. Lett.* **95,** 190405 (2005).
54. Vaucher, B., Nunnenkamp, A. & Jaksch, D. Creation of resilient entangled states and a resource for measurement-based quantum computation with optical superlattices. Preprint at ⟨http://arxiv.org/abs/0710.5099⟩ (2007).
55. Deutsch, D. Quantum computational networks. *Proc. R. Soc. Lond. A* **425,** 73–90 (1989).
56. Yao, A. in *Proc. 34th Annu. Symp. Found. Comput. Sci.* 352–361 (IEEE Computer Soc., Los Alamitos, 1993).
57. Barenco, A. *et al.* Elementary gates for quantum computation. *Phys. Rev. A* **52,** 3457–3467 (1995).
58. Farhi, E. *et al.* A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem. *Science* **292,** 472–476 (2001).
59. Deutsch, D. Quantum-theory, the Church–Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A* **400,** 97–117 (1985).
60. Bernstein, E. & Vazirani, U. in *Proc. 25th Annu. ACM Symp. Theor. Comput.* 11–20 (ACM Press, New York, 1993).
61. Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* **402,** 390–393 (1999).
62. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409,** 46–52 (2001).
63. Nielsen, M. A. Quantum computation by measurement and quantum memory. *Phys. Lett. A* **308,** 96–100 (2003).
64. Raussendorf, R. & Briegel, H. J. A One-way quantum computer. *Phys. Rev. Lett.* **86,** 5188–5191 (2001).
65. Raussendorf, R. & Briegel, H. J. Computational model underlying the one-way quantum computer. *Quant. Info. Comput.* **2,** 443–486 (2002).
66. Walther, P. *et al.* Experimental one-way quantum computing. *Nature* **434,** 169–176 (2005).
67. Kiesel, N. *et al.* Experimental analysis of a four-qubit photon cluster state. *Phys. Rev. Lett.* **95,** 210502 (2005).
68. Gross, D., Eisert, J., Schuch, N. & Perez-Garcia, D. Measurement-based quantum computation beyond the one-way model. *Phys. Rev. A* **76,** 052315 (2007).
69. Van den Nest, M., Miyake, A., Dür, W. & Briegel, H. J. Universal resources for measurement-based quantum computation. *Phys. Rev. Lett.* **97,** 150504 (2006).
70. Van den Nest, M., Dür, W., Miyake, A. & Briegel, H. J. Fundamentals of universality in one-way quantum computation. *New J. Phys.* **9,** 204 (2007).
71. Gross, D. & Eisert, J. Novel schemes for measurement-based quantum computation. *Phys. Rev. Lett.* **98,** 220503 (2007).
72. Browne, D. E. *et al.* Phase transition of computational power in the resource states for one-way quantum computation. Preprint at ⟨http://arxiv.org/abs/0709.1729⟩ (2007).
73. Verstraete, F. & Cirac, J. I. Valence-bond states for quantum computation. *Phys. Rev. A* **70,** 060302 (2004).
74. Aliferis, P. & Leung, D. W. Computation by measurements: a unifying picture. *Phys. Rev. A* **70,** 062314 (2004).
75. Childs, A. M., Leung, D. W. & Nielsen, M. A. Unified derivations of measurement-based schemes for quantum computation. *Phys. Rev. A* **71,** 032318 (2005).
76. Jorrand, P. & Perdrix, S. Unifying quantum computation with projective measurements only and one-way quantum computation. Preprint at ⟨http://arxiv.org/abs/quant-ph/0404125⟩ (2004).
77. Shor, P. W. in *Proc. 37th Annu. Symp. Found. Comput. Sci.* 56–65 (IEEE Computer Soc., Los Alamitos, 1996).
78. Aharonov, D. & Ben-Or, M. in *Proc. 29th Annu. ACM Symp. Theor. Comput.* 176–188 (ACM Press, New York, 1997).
79. Gottesman, D. *Stabilizer Codes and Quantum Error Correction*. PhD thesis, California Inst. Technol. (1997).
80. Knill, E., Laflamme, R. & Zurek, W. H. Resilient quantum computation: error models and thresholds. *Proc. R. Soc. Lond. A* **454,** 365–384 (1998).
81. Kitaev, A. Y. Fault-tolerant quantum computation by anyons. *Ann. Phys. (NY)* **303,** 2–30 (2003).
82. Raussendorf, R. & Harrington, J. Fault-tolerant quantum computation with high threshold in two dimensions. *Phys. Rev. Lett.* **98,** 190504 (2007).
83. Hofstetter, W., Cirac, J. I., Zoller, P., Demler, E. & Lukin, M. D. High-temperature superfluidity of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **89,** 220407 (2002).
84. Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78,** 17–85 (2006).
85. Anderson, P. W. The resonating valence bond state in La_2CuO_4 and superconductivity. *Science* **235,** 1196–1198 (1987).
86. Werner, F., Parcollet, O., Georges, A. & Hassan, S. R. Interaction-induced adiabatic cooling and antiferromagnetism of cold fermions in optical lattices. *Phys. Rev. Lett.* **95,** 056401 (2005).

Acknowledgements I thank H. Briegel for discussions, and the German Research Foundation (DFG), the European Union (through the OLAQUI and SCALA projects) and the Air Force Office of Scientific Research (AFOSR) for support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (bloch@uni-mainz.de).

The quantum internet

H. J. Kimble[1]

Quantum networks provide opportunities and challenges across a range of intellectual and technical frontiers, including quantum computation, communication and metrology. The realization of quantum networks composed of many nodes and channels requires new scientific capabilities for generating and characterizing quantum coherence and entanglement. Fundamental to this endeavour are quantum interconnects, which convert quantum states from one physical system to those of another in a reversible manner. Such quantum connectivity in networks can be achieved by the optical interactions of single photons and atoms, allowing the distribution of entanglement across the network and the teleportation of quantum states between nodes.

In the past two decades, a broad range of fundamental discoveries have been made in the field of quantum information science, from a quantum algorithm that places public-key cryptography at risk to a protocol for the teleportation of quantum states[1]. This union of quantum mechanics and information science has allowed great advances in the understanding of the quantum world and in the ability to control coherently individual quantum systems[2]. Unique ways in which quantum systems process and distribute information have been identified, and powerful new perspectives for understanding the complexity and subtleties of quantum dynamical phenomena have emerged.

In the broad context of quantum information science, quantum networks have an important role, both for the formal analysis and the physical implementation of quantum computing, communication and metrology[2–5]. A notional quantum network based on proposals in refs 4, 6 is shown in Fig. 1a. Quantum information is generated, processed and stored locally in quantum nodes. These nodes are linked by quantum channels, which transport quantum states from site to site with high fidelity and distribute entanglement across the entire network. As an extension of this idea, a 'quantum internet' can be envisaged; with only moderate processing capabilities, such an internet could accomplish tasks that are impossible in the realm of classical physics, including the distribution of 'quantum software'[7].

Apart from the advantages that might be gained from a particular algorithm, there is an important advantage in using quantum connectivity, as opposed to classical connectivity, between nodes. A network of quantum nodes that is linked by classical channels and comprises k nodes each with n quantum bits (qubits) has a state space of dimension $k2^n$, whereas a fully quantum network has an exponentially larger state space, 2^{kn}. Quantum connectivity also provides a potentially powerful means to overcome size-scaling and error-correlation problems that would limit the size of machines for quantum processing[8]. At any stage in the development of quantum technologies, there will be a largest size attainable for the state space of individual quantum processing units, and it will be possible to surpass this size by linking such units together into a fully quantum network.

A different perspective of a quantum network is to view the nodes as components of a physical system that interact by way of the quantum channels. In this case, the underlying physical processes used for quantum network protocols are adapted to simulate the evolution of quantum many-body systems[9]. For example, atoms that are localized at separate nodes can have effective spin–spin interactions catalysed by single-photon pulses that travel along the channels between the nodes[10]. This 'quantum wiring' of the network allows a wide range for the effective hamiltonian and for the topology of the resultant 'lattice'. Moreover, from this perspective, the extension of entanglement across quantum networks can be related to the classical problem of percolation[11].

These exciting opportunities provide the motivation to examine research related to the physical processes for translating the abstract illustration in Fig. 1a into reality. Such considerations are timely because scientific capabilities are now passing the threshold from a learning phase with individual systems and advancing into a domain of rudimentary functionality for quantum nodes connected by quantum channels.

In this review, I convey some basic principles for the physical implementation of quantum networks, with the aim of stimulating the involvement of a larger community in this endeavour, including in systems-level studies. I focus on current efforts to harness optical processes at the level of single photons and atoms for the transportation of quantum states reliably across complex quantum networks.

Two important research areas are strong coupling of single photons and atoms in the setting of cavity quantum electrodynamics (QED)[12] and quantum information processing with atomic ensembles[13], for which crucial elements are long-lived quantum memories provided by the atomic system and efficient, quantum interfaces between light and matter. Many other physical systems are also being investigated and are discussed elsewhere (ref. 2 and websites for the Quantum Computation Roadmap (http://qist.lanl.gov/qcomp_map.shtml), the SCALA Integrated Project (http://www.scala-ip.org/public) and Qubit Applications (http://www.qubitapplications.com)).

A quantum interface between light and matter

The main scientific challenge in the quest to distribute quantum states across a quantum network is to attain coherent control over the interactions of light and matter at the single-photon level. In contrast to atoms and electrons, which have relatively large long-range interactions for their spin and charge degrees of freedom, individual photons typically have interaction cross-sections that are orders of magnitude too small for non-trivial dynamics when coupled to single degrees of freedom for a material system.

The optical physics community began to address this issue in the 1990s, with the development of theoretical protocols for the coherent transfer of quantum states between atoms and photons in the setting of cavity QED[6,14,15]. Other important advances have been made in the past

[1]Norman Bridge Laboratory of Physics 12–33, California Institute of Technology, Pasadena, California 91125, USA.

decade[2,4], including with atomic ensembles[13,16]. The reversible mapping of quantum states between light and matter provides the basis for quantum-optical interconnects and is a fundamental primitive (building block) for quantum networks. Although the original schemes for such interconnects are sensitive to experimental imperfections, a complete set of theoretical protocols has subsequently been developed for the robust distribution of quantum information over quantum networks, including, importantly, the quantum repeater[4,17] and scalable quantum networks with atomic ensembles[13].

A generic quantum interface between light and matter is depicted in Fig. 1b. This interface is described by the interaction hamiltonian $H_{int}(t)$, where for typical states $H_{int}(t)\rangle \approx \hbar\chi(t)$, with $\hbar$ being $h/2\pi$ (where h is Planck's constant) and $\chi(t)$ being the time-dependent coupling strength between the internal material system and the electromagnetic field. Desirable properties for a quantum interface include that $\chi(t)$ should be 'user controlled' for the clocking of states to and from the quantum memory (for example, by using an auxiliary laser), that the physical processes used should be robust in the face of imperfections (for example, by using adiabatic transfer) and that mistakes should be efficiently detected and fixed (for example, with quantum error correction). In qualitative terms, the rate κ, which characterizes the bandwidth of the input–output channel, should be large compared with the rate γ, which characterizes parasitic losses, and both of these rates should be small compared with the rate of coherent coupling χ.

Examples of physical systems for realizing a quantum interface and distributing coherence and entanglement between nodes are shown in Fig. 1c, d. In the first example (Fig. 1c), single atoms are trapped in optical cavities at nodes A and B, which are linked by an optical fibre. External fields control the transfer of the quantum state $|\Psi\rangle$ stored in the atom at node A to the atom at node B by way of photons that propagate from node A to node B[6,18]. In the second example (Fig. 1d), a single-photon pulse that is generated at node A is coherently split into two

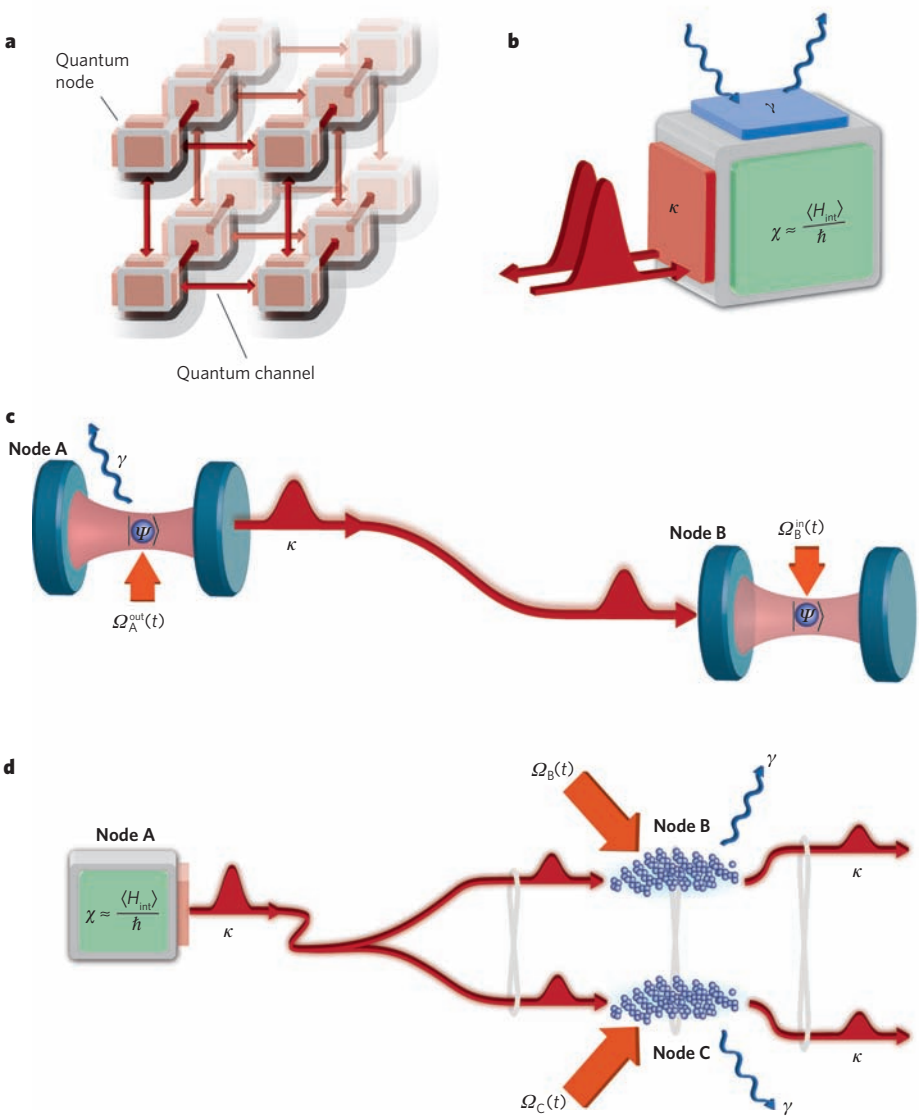


Figure 1 | Quantum networks. a, Shown is a notional quantum network composed of quantum nodes for processing and storing quantum states and quantum channels for distributing quantum information. Alternatively, such a network can be viewed as a strongly correlated many-particle system. **b**, The quantum interface between matter (coloured cube) and light (red curves) is depicted. Coherent interactions in the node are characterized by the rate χ; coupling between the node and photons in the external channel occurs at the rate κ; and parasitic losses occur at the rate γ. **c**, Quantum state transfer and entanglement distribution from node A to node B is shown in the setting of cavity quantum electrodynamics (QED)[6]. At node A, a pulse of the control field $\Omega_A^{out}(t)$ causes the transformation of atomic state $|\Psi\rangle$ into the state of a propagating optical field (that is, into a flying photon). At node B, the pulse $\Omega_B^{in}(t)$ is applied to map the state of the flying photon into an atom in the cavity, thereby realizing the transfer of the state $|\Psi\rangle$ from node A to node B (ref. 18). **d**, The distribution of entanglement by using ensembles of a large number of atoms is shown[13]. A single-photon pulse at node A is coherently split into two entangled components that propagate to node B and node C and then are coherently mapped by the control fields $\Omega_{B,C}^{in}(t)$ into a state that is entangled between collective excitations in each ensemble at node B and node C. At later times, components of the entangled state can be retrieved from the quantum memories by separate control fields, $\Omega_{B,C}^{out}(t)$ (ref. 19). $H_{int}(t)$, interaction hamiltonian; $\hbar$, $h/2\pi$ (where h is Planck's constant).

components and propagates to nodes B and C, where the entangled photon state is coherently mapped into an entangled state between collective excitations at each of the two nodes[13,19]. Subsequent read-out of entanglement from the memories at node B and/or node C as photon pulses is implemented at the 'push of a button'.

Cavity QED

At the forefront of efforts to achieve strong, coherent interactions between light and matter has been the study of cavity QED[20]. In both the optical[12,21] and the microwave[22–25] domains, strong coupling of single atoms and photons has been achieved by using electromagnetic resonators of small mode volume (or cavity volume) V_m with quality factors $Q \approx 10^7$–10^{11}. Extensions of cavity QED to other systems[26] include quantum dots coupled to micropillars and photonic bandgap cavities[27], and Cooper pairs interacting with superconducting resonators (that is, circuit QED; see ref. 28 for a review).

Physical basis of strong coupling

Depicted in Fig. 2a is a single atom that is located in an optical resonator and for which strong coupling to a photon requires that a single intracavity photon creates a 'large' electric field. Stated more quantitatively, if the coupling frequency of one atom to a single mode of an optical resonator is g (that is, $2g$ is the one-photon Rabi frequency), then

$$g = \sqrt{\frac{|\boldsymbol{\varepsilon}\cdot\boldsymbol{\mu}_0|^2\,\omega_C}{2\hbar\varepsilon_0 V_m}} \qquad (1)$$

where $\boldsymbol{\mu}_0$ is the transition dipole moment between the relevant atomic states (with transition frequency ω_A), and $\omega_C \approx \omega_A$ is the resonant frequency of the cavity field, with polarization vector $\boldsymbol{\varepsilon}$. Experiments in cavity QED explore strong coupling with $g >> (\gamma, \kappa)$, where γ is the atomic decay rate to modes other than the cavity mode and κ is the decay rate of the cavity mode itself. Expressed in the language of traditional optical physics, the number of photons required to saturate the intracavity atom is $n_0 \approx \gamma^2/g^2$, and the number of atoms required to have an appreciable effect on the intracavity field is $N_0 \approx \kappa\gamma/g^2$. Strong coupling in cavity QED moves beyond traditional optical physics, for which $(n_0, N_0) >> 1$, to explore a qualitatively new regime with $(n_0, N_0) << 1$ (ref. 12).

In the past three decades, a variety of approaches have been used to achieve strong coupling in cavity QED[12,20–25]. In the optical domain, a route to strong coupling is the use of high-finesse optical resonators ($F \approx 10^5$–10^6) and atomic transitions with a large $\boldsymbol{\mu}_0$ (that is, oscillator strengths near unity). Progress along this path is illustrated in Fig. 2c, with research now far into the domain $(n_0, N_0) << 1$.

As the cavity volume V_m is reduced to increase g (equation (1)), the requirement for atomic localization becomes more stringent. Not surprisingly, efforts to trap and localize atoms in high-finesse optical cavities in a regime of strong coupling have been central to studies of cavity QED in the past decade, and the initial demonstration was in 1999 (ref. 29). Subsequent advances include extending the time for which an atom is trapped to 10 s (refs 30, 31); see ref. 32 for a review. Quantum control over both internal degrees of freedom (that is, the atomic dipole and the cavity field) and external degrees of freedom (that is, atomic motion) has now been achieved for a strongly coupled atom–cavity system[33]. And an exciting prospect is cavity QED with single trapped ions, for which the boundary for strong coupling has been reached[34].

Coherence and entanglement in cavity QED

Applying these advances to quantum networks has allowed single photons to be generated 'on demand' (Box 1). Through strong coupling of the cavity field to an atomic transition, an external control field $\Omega(t)$ transfers one photon into the cavity mode and then to free space by way of the cavity output mirror, leading to a single-photon pulse $|\phi_1(t)\rangle$ as a collimated beam. The temporal structure (both amplitude and phase) of the resultant 'flying photon' $|\phi_1(t)\rangle$ can be tailored by way of the control field $\Omega(t)$ (refs 6, 35), with the spatial structure of the wave packet being set by the cavity mode.

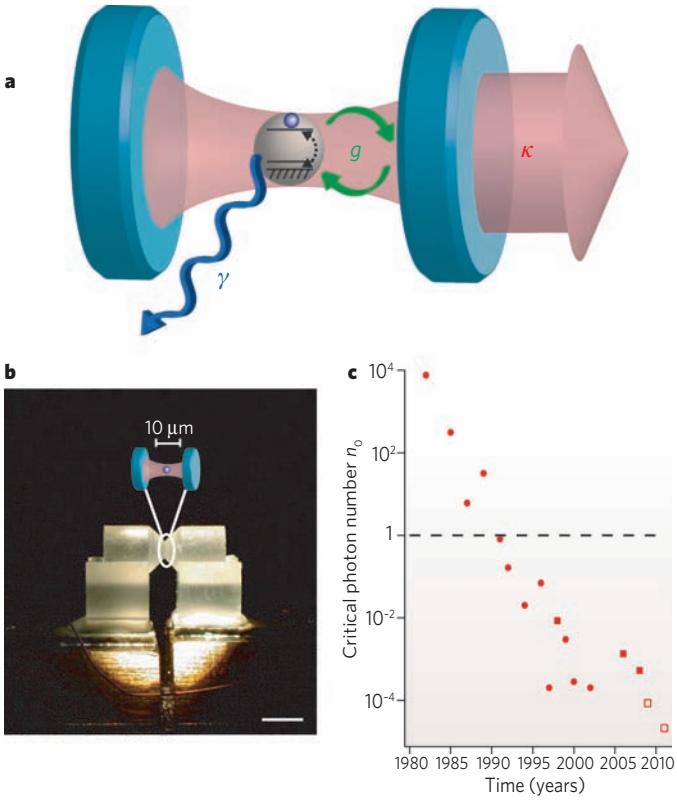
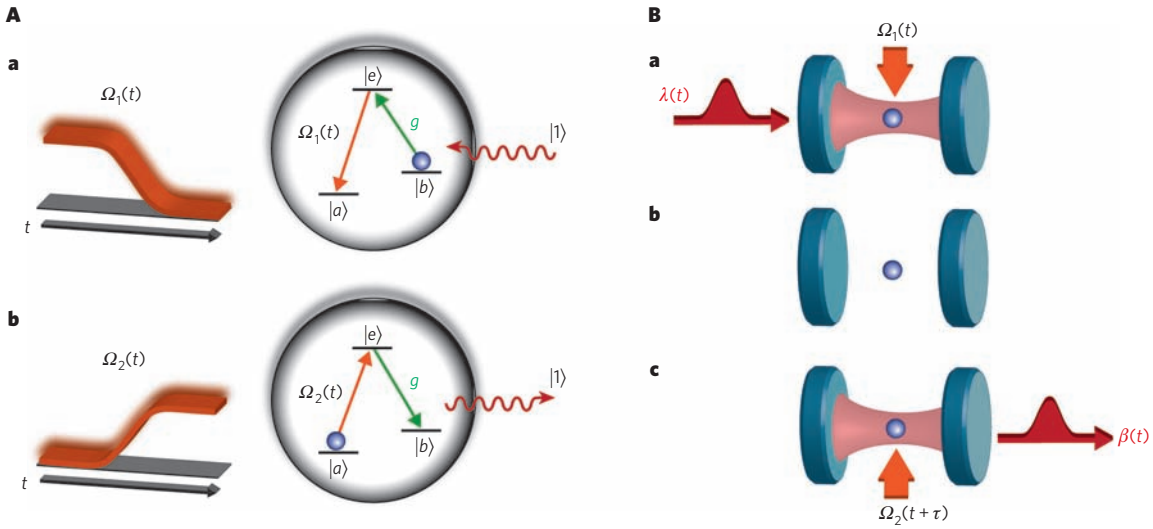


Figure 2 | Elements of cavity QED. a, Shown is a simple schematic of an atom–cavity system depicting the three governing rates (g, κ, γ) in cavity QED, where $g \approx \chi$ in Fig. 1. Coherent exchange of excitation between the atom and the cavity field proceeds at rate g, as indicated by the dashed arrow for the atom and the green arrows for the cavity field. **b**, A photograph of two mirror substrates that form the Fabry–Pérot cavity, which is also shown schematically. The cavity length $l = 10\,\mu m$, waist $w_0 = 12\,\mu m$ transverse to the cavity axis, and finesse $F \approx 5 \times 10^5$. The supporting structure allows active servo control of the cavity length to $\delta l \approx 10^{-14}$ m (ref. 12). Scale bar, 3 mm. **c**, The reduction in the critical photon number n_0 over time is shown for a series of experiments in cavity QED that were carried out by the Caltech Quantum Optics Group. These experiments involved either spherical-mirror Fabry–Pérot cavities (circles) or the whispering-gallery modes of monolithic SiO_2 resonators (squares). The data points shown for 2006 and 2008 are for a microtoroidal SiO_2 resonator[75,76]; those for 2009 and 2011 (open squares) are projections for this type of resonator[77].

Several experiments have confirmed the essential aspects of this process for the deterministic generation of single photons[30,34,36]. Significantly, in the ideal (adiabatic) limit, the excited state $|e\rangle$ of the atom is not populated because of the use of a 'dark state' protocol[37]. By deterministically generating a bit stream of single-photon pulses from single trapped atoms, these experiments are a first step in the development of quantum networks based on flying photons.

Compared with the generation of single photons by a variety of other systems[38], one of the distinguishing aspects of the dark-state protocol (Box 1) is that it should be reversible. That is, a photon that is emitted from a system A should be able to be efficiently transferred to another system B by applying the time-reversed (and suitably delayed) field $\Omega(t)$ to system B (Fig. 1c). Such an advance was made[18] by implementing the reversible mapping of a coherent optical field to and from internal states of a single trapped caesium atom. Although this experiment was imperfect, it provides the initial verification of the fundamental primitive on which the protocol for the physical implementation of quantum networks in ref. 6 is based (an important theoretical protocol that has been adapted to many theoretical and experimental settings).

Box 1 | Mapping quantum states between atoms and photons



Reversible transfer of a state between light and a single trapped atom can be achieved through the mappings $|b\rangle|1\rangle \rightarrow |a\rangle|0\rangle$ and $|a\rangle|0\rangle \rightarrow |b\rangle|1\rangle$ for the coherent absorption and emission of single photons (in panel **A**, **a** and **b** of the figure, respectively)[18]. In this case, $|a\rangle$ and $|b\rangle$ represent internal states of the atom with long-lived coherence (for example, atomic hyperfine states in the $6S_{1/2}$, $F = 3$ and $F = 4$ manifolds of atomic caesium), and $|0\rangle$ and $|1\rangle$ are Fock states of the photons in the intracavity field with $n = 0$ and $n = 1$ excitations, respectively. The transition between $|b\rangle$ and $|e\rangle$ is strongly coupled to a mode of an optical cavity with interaction energy $\hbar g$, where g (in green) is the coherent coupling rate of the atom and the photon. In this simple setting, the interaction hamiltonian for atom and cavity field has a dark state $|D\rangle$ (that is, there is no excited state component $|e\rangle$)[37], as given by $|D\rangle = \cos\theta|a\rangle|0\rangle + \sin\theta|b\rangle|1\rangle$, where

$$\cos\theta = \left[1 + \frac{\Omega^2(t)}{g^2}\right]^{-1/2} \quad (1)$$

with $\Omega(t)$ as a classical control field[14]. For $\Omega(t = 0) = 0$, then $|D\rangle = |a\rangle|0\rangle$. By contrast, for $\Omega(t \rightarrow \infty) >> g$, $|D\rangle \rightarrow |b\rangle|1\rangle$.

Panel **A**, **a** of the figure shows that by adiabatically ramping a control field $\Omega_1(t) >> g$ from on to off over a time Δt that is slow compared with $1/g$, the atomic state is mapped from $|b\rangle$ to $|a\rangle$ with the accompanying coherent absorption of one intracavity photon. Conversely, in panel **A**, **b** of the figure, by turning a control field $\Omega_2(t)$ from off to on, the atomic state is mapped from $|a\rangle$ to $|b\rangle$ with the transfer of one photon into the cavity mode.

These two processes can be combined to achieve the coherent transfer of the state of a propagating optical field $\lambda(t) = |\phi_{\text{field}}(t)\rangle$ into and out of a quantum memory formed by the atomic states $|a\rangle$ and $|b\rangle$ (ref. 18; figure, panel **B**). In the ideal case, the mapping is specified by $|\phi_{\text{field}}(t)\rangle|b\rangle \rightarrow |0\rangle(c_1|a\rangle + c_0|b\rangle) \ldots$ (storage) $\ldots |0\rangle(c_1|a\rangle + c_0|b\rangle) \rightarrow |\phi_{\text{field}}(t+\tau)\rangle|b\rangle$, where the field state is taken to be a coherent superposition of zero (c_0) and one (c_1) photon, $|\phi_{\text{field}}(t)\rangle = E(t)[c_0|0\rangle_{\text{field}} + c_1|1\rangle_{\text{field}}]$. $E(t)$ is the envelope of the field external to the cavity, with $\int|E(t)|^2dt = 1$; $t+\tau$ is a user-selected time (discussed below). Given timing information for the incoming field $|\phi_{\text{field}}(t)\rangle$, the first step in this process (figure, panel **B**, **a**) is accomplished by adiabatically ramping the control field $\Omega_1(t)$ from on to off, as in **A**, **a**. After this step, the internal states of the atom provide a long-lived quantum memory (figure, panel **B**, **b**). At a user-selected later time $t+\tau$, the final step is initiated (figure, panel **B**, **c**) by turning $\Omega_2(t+\tau)$ from off to on (as in **A**, **b**), thereby coherently mapping the atomic state $c_1|a\rangle + c_0|b\rangle$ back to the 'flying' field state $\beta(t) = |\phi_{\text{field}}(t+\tau)\rangle$.

The adiabatic transfer of quantum states (as described in Box 1, as well as related possibilities[10,35]) relies on strong coupling between an atom and a single polarization of the intracavity field. However, by extending the ideas in Box 1 to the two polarization eigenmodes of the cavity for given transverse and longitudinal mode orders, it is possible to generate entanglement between the internal states of the atom and the polarization state of a coherently generated photon[39–41]. An initial control field $\Omega_1(t_1)$ results in entanglement between internal states of the atom b, $|b_\pm\rangle$, and the polarization state of a flying photon $|\phi^{\pm}_{\text{field}}(t_1)\rangle$ that is coherently generated by the coupled atom–cavity system. Applying a second control field $\Omega_2(t_2)$ returns the atom to its initial (unentangled) state while generating a second flying photon $|\xi^{\pm}_{\text{field}}(t_2)\rangle$, thereby leading to entanglement between the polarizations of the fields, $\phi^{\pm}_{\text{field}}$ and $\xi^{\pm}_{\text{field}}$, emitted at times t_1 and t_2.

Such a sequence of operations has been applied to single rubidium atoms falling through a high-finesse optical cavity[21]. In this study, entangled photons were generated with a time separation $\tau = t_2 - t_1$ limited by the atomic transit time. Although the atoms arrived randomly into the cavity mode in this case, the protocol itself is intrinsically deterministic. With trapped atoms, it will be possible to generate entangled states at user selected times (t_1, t_2) at the 'push of a button.' Moreover, the scheme is inherently reversible, so the entanglement between atom and field can be used to distribute entanglement to a second atom–cavity system in a network.

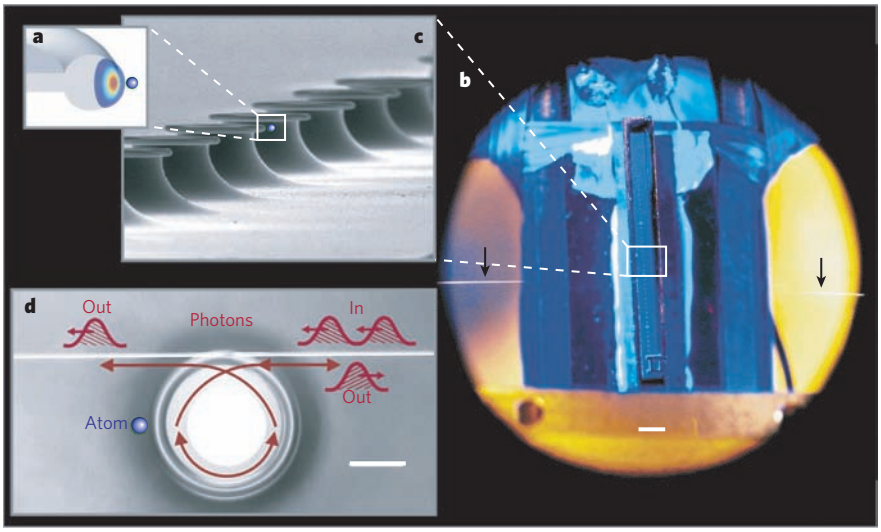
In a broader context, important advances have been made in the generation and transfer of quantum states in other physical systems, including quantum dots[42] and circuits[28] coupled to cavities.

With the maturation of experimental capabilities in cavity QED that is now evident, many previously developed theoretical protocols will become possible. These include the sequential generation of entangled multiqubit states[43], the teleportation of atomic states from one node to another[15], photonic quantum computation by way of photon–photon interactions at the nodes[35] and reversible mapping of quantum states of atomic motion to and from light[44]. Clearly, new technical capabilities beyond conventional (Fabry–Pérot) cavities will be required to facilitate such scientific investigations; several candidate systems are discussed in Box 2.

Quantum networks with atomic ensembles

An area of considerable research activity in the quest to distribute coherence and entanglement across quantum networks has been the interaction of light with atomic ensembles that consist of a large collection of identical atoms. For the regime of continuous variables, entanglement has been achieved between two atomic ensembles, each of which consists of ~10^{12} atoms[45], and the quantum teleportation of light to matter has been demonstrated by mapping coherent optical states to the collective spin states of an atomic memory[46]. Further research of the continuous variables regime is reviewed elsewhere[47]. Here I focus, instead, on the regime of discrete variables, with photons and atomic excitations considered one by one.

Box 2 | A new paradigm for cavity QED



To build large-scale quantum networks[4,6], many quantum nodes will need to be interconnected over quantum channels. Because conventional (Fabry–Pérot) configurations are ill suited for this purpose, there have been efforts to develop alternative microcavity systems[26], both for single atoms[75,76,78] and for atom-like systems (such as nitrogen-vacancy centres in diamond[79]). A quantitative comparison of candidate systems is provided in ref. 77.

A remarkable resonator for this purpose is the microtoroidal cavity that is formed from fused SiO_2 (refs 80,81) (shown in the figure). Such a resonator supports a whispering-gallery mode[82] circulating around the outer circumference of the toroid (shown in cross-section in grey, in panel **a** of the figure), with an evanescent field external to the resonator. The intensity of the resonator mode is indicated by the coloured contours. Because of the small mode volume V_m and large quality factor Q, an atom (blue) interacting with the evanescent field of a whispering-gallery mode can be far into the regime of strong coupling, with projected values for the critical photon n_0 and atom N_0 numbers ($n_0 \approx 2 \times 10^{-5}$ and $N_0 \approx 10^{-6}$)[77] that are significantly greater than current[12] and projected[77] values for cavity QED with Fabry–Pérot cavities (Fig. 2c).

Pioneering fabrication techniques[80,81] lend themselves to the integration of many microtoroidal resonators to form optical networks, as illustrated in panel **b** and **c** of the figure. Panel **b** shows a photograph of a silicon chip with a linear array of microtoroidal resonators within an ultrahigh-vacuum apparatus[76]. The toroids appear as small scattering centres on a silicon chip that runs vertically down the centre of the picture. Black arrows indicate a horizontal SiO_2 fibre taper for coupling light to and from one resonator. Scale bar, 2 mm. Panel **c** is a scanning electron micrograph of an array of microtoroidal resonators (a magnification of the region bounded by the white box in panel **b**), showing toroids of fused SiO_2 on silicon supports[80].

These resonators have the capability for input–output coupling with small parasitic loss[81] for the configuration shown in panel **d** (scale bar, 10 μm), which is a micrograph of an individual toroid and fibre taper from panel **b**[76]. $Q = 4 \times 10^8$ has been realized at $\lambda = 1{,}550$ nm, and $Q \approx 10^8$ at $\lambda = 850$ nm, with good prospects for improvement to $Q \approx 10^{10}$ (ref. 77). For these parameters, the efficiency ε for coupling quantum fields into and out of the resonator could approach $\varepsilon \approx 0.99$–0.999 while remaining firmly in the regime of strong coupling[77]. Such high efficiency is crucial for the realization of complex quantum networks, including for distributing and processing quantum information[4,6,35] and for investigating the association between quantum many-body systems and quantum networks[9,11].

The initial step in this quest to realize a quantum network was the demonstration of strong coupling between individual atoms and the field of a microtoroidal resonator[75]. More recently, non-classical fields have been generated from the interaction of single atoms with a microtoroidal resonator by way of a 'photon turnstile', for which a single atom dynamically regulates the transport of photons one by one through the microtoroidal resonator[76] (figure, panel **d**). Only single photons can be transmitted in the forward direction (from right to left in the figure), with excess photons $n > 1$ dynamically rerouted to the backward direction.

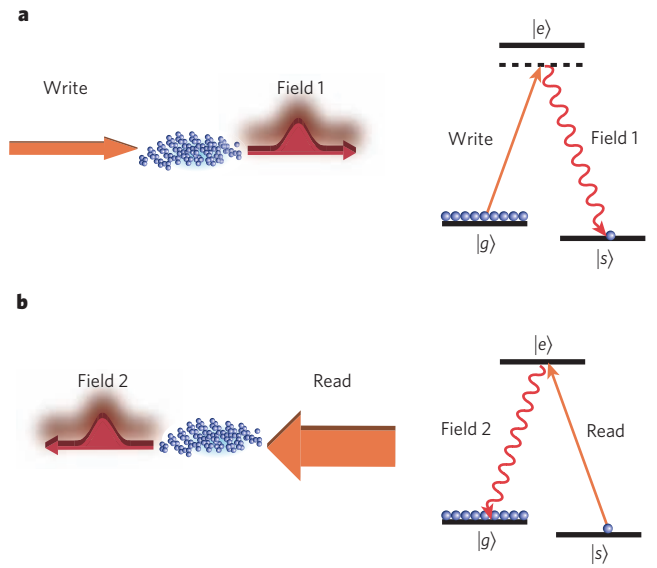
Writing and reading collective spin excitations

Research on discrete quantum variables is based on the remarkable theoretical protocol described in ref. 13, in which Luming Duan, Mikhail Lukin, Juan Ignacio Cirac and Peter Zoller presented a realistic scheme for entanglement distribution by way of a quantum-repeater architecture[4,17]. Fundamental to this protocol, which is known as the DLCZ protocol, is the generation and retrieval of single 'spin' excitations within an ensemble of a large number of atoms[48] (Box 3). Together with photoelectric detection of field 1, a laser pulse ('write' pulse) creates a single excitation $|1_a\rangle$ that is stored collectively within the atomic ensemble. At a later time, a second laser pulse ('read' pulse) deterministically converts excitation stored within the atomic memory in the state $|1_a\rangle$ into a propagating field, denoted field 2.

The basic processes illustrated in Box 3 can be extended to create an entangled pair of ensembles, L and R (ref. 13; Fig. 3a). The entangled state is generated in a probabilistic but heralded[49] manner from quantum interference in the measurement process. That is, detection of a photon from one atomic ensemble or the other in an indistinguishable manner results in an entangled state with one collective spin excitation shared coherently between the ensembles. In the ideal case, and to lowest-order probability, a photoelectric detection event at either of the two detectors projects the ensembles into the entangled state $|\Psi_{L,R}\rangle = \frac{1}{\sqrt{2}}(|0_a\rangle_L|1_a\rangle_R \pm e^{i\eta_1}|1_a\rangle_L|0_a\rangle_R)$, with the sign (+ or −) set by whether detector 1 or detector 2 records the event. The phase η_1 is determined by the difference between the phase shifts along the two channels, $\eta_1 = \beta_L - \beta_R$ (ref. 49), which must be stable. Any given trial with a 'write' pulse is unlikely to produce a detection event at either detector, and such failed trials require the system to be reinitialized. However, a photoelectric detection event at either detector unambiguously heralds the creation of the entangled state. Limited by the coherence time between the metastable lower atomic states $|g\rangle_i$ and $|s\rangle_i$ for all atoms $i = 1, 2, \ldots, N_a$ within the ensemble (ref. 50; Box 3), this entangled state is stored in the quantum memory provided by the ensembles and is available 'on demand' for subsequent tasks, such as entanglement connection[13,51].

Although the above description is for an ideal case and neglects higher-order terms, the DLCZ protocol is designed to be resilient to

Box 3 | Writing and reading single atomic excitations



The DLCZ protocol[13] is based on ensembles of N_a identical atoms (blue) with a Λ-level configuration, as shown in the figure. The metastable lower states $|g\rangle$ and $|s\rangle$ can be, for example, atomic hyperfine states of the electronic ground level to ensure a long lifetime for coherence. All atoms are initially prepared in state $|g\rangle$ with no excitation (figure, panel **a**), namely $|0_a\rangle \equiv \otimes_i^{N_a}|g\rangle_i$, and a weak off-resonant 'write' pulse is then sent through the ensemble. This results in a small probability of amplitude $\sqrt{p}$ that one of the N_a atoms will be transferred from $|g\rangle$ to $|s\rangle$ and will emit a photon into the forward-scattered optical mode (designated field 1) with a frequency and/or polarization distinct from the write field.

For small excitation probability $p<<1$, in most cases nothing happens as a result of the writing pulse, so the resultant state $|\phi_{a,1}\rangle$ for the atomic ensemble and field 1 in the ideal case is given by

$$|\phi_{a,1}\rangle \quad |0_a\rangle|0_1\rangle + e^{i\beta}\sqrt{p}|1_a\rangle|1_1\rangle + O(p) \qquad (1)$$

where $|n_1\rangle$ is the state of the forward-propagating field 1 with n_1 photons ($n_1 = 0$ or 1), the phase β is determined by the propagation phases of the write pulse and field 1, and $O(p)$ denotes of order p. The atomic state $|1_a\rangle$ in equation (1) (above) is a collective (entangled) state with one excitation shared symmetrically between the N_a atoms (that is, one 'spin flip' from $|g\rangle$ to $|s\rangle$), where in the ideal case[13]

$$|1_a\rangle = \frac{1}{\sqrt{N_a}}\sum_{i=1}^{N_a}|g\rangle_i \cdots |s\rangle_i \cdots |g\rangle_{N_a} \qquad (2)$$

Field 1 is directed to a single-photon detector, where a detection event is recorded with probability p. Such an event for field 1 heralds that a single excitation (or spin flip from $|g\rangle$ to $|s\rangle$) has been created and stored in the atomic ensemble in the state $|1_a\rangle$ with high probability. Higher-order processes with multiple atomic and field 1 excitations are also possible and ideally occur, to lowest order, with probability p^2.

After a user-defined delay (subject to the finite lifetime of the quantum memory), the collective atomic excitation $|1_a\rangle$ can be efficiently converted to a propagating beam (designated field 2) by way of a strong 'read' pulse (figure, panel **b**), where in the ideal case there is a one-to-one transformation of atomic excitation to field excitation, $|1_a\rangle$ to $|1_2\rangle$. In the case of resonance with the transition from $|s\rangle$ to $|e\rangle$, the reading process utilizes the phenomenon of electromagnetically induced transparency[16,66].

important sources of imperfections, including losses in propagation and detection, and detector dark counts. Indeed, the scheme functions with 'built-in entanglement purification'[13] and enables entanglement to be extended beyond the separation of two ensembles in an efficient and scalable manner. Theoretical extensions[52,53] of the DLCZ protocol have examined related network architectures for optimizing scalability in view of laboratory capabilities (discussed below).

Coherence and entanglement with atomic ensembles

The initial, enabling, steps in the implementation of the DLCZ protocol were observations of quantum correlations both for single photon pairs[54,55] and for a large number of photons (10^3–10^4) (ref. 56) generated in the collective emission from atomic ensembles. Single photons were generated by the efficient mapping of stored collective atomic excitation to propagating wave packets for field 2 (refs 57–61; Box 3). Conditional read-out efficiencies of 50% in free space[58] and 84% in a cavity[62] were realized for state transfer from a single collective 'spin' excitation stored in the atomic ensemble to a single photon for field 2.

With these capabilities for coherent control of collective atomic emission, heralded entanglement between ensembles separated by 3 m was achieved in 2005 (ref. 49). More recent work has led to the inference that the concurrence C (ref. 63) of entanglement stored between the two ensembles in Fig. 3 is $C = 0.9\pm0.3$ (ref. 50), with the associated density matrix shown in Fig. 3b.

The DLCZ protocol is based on a quantum-repeater architecture involving independent operations on parallel chains of quantum systems[13], with scalability relying crucially on conditional control of quantum states stored in remote quantum memories[64]. The experiment shown in Fig. 3c took an important step towards this goal by achieving the minimal functionality required for scalable quantum networks[65].

Apart from the DLCZ protocol, which involves measurement-induced entanglement, it is also possible to achieve deterministic mapping of quantum states of light into and out of atomic ensembles by using electromagnetically induced transparency[16,66]. Pioneering work[67,68] demonstrated the storage and retrieval of classical pulses to and from an atomic ensemble. This work was then extended into the quantum regime of single photons[69,70]. Entanglement between two ensembles coupled to a cavity mode was achieved by adiabatic transfer of excitation[71], thereby providing a means for on-demand entanglement. In addition, the reversible mapping of photonic entanglement into and out of pairs of quantum memories has been achieved[19] by an electromagnetically-induced-transparency process, which should assist the distribution of entanglement over quantum networks (Fig. 1d).

Contemporary with this work on heralded and deterministic entanglement, a variety of experiments based on entanglement as a postdiction have been carried out[72] (that is, for cases in which a physical state is not available for use in a scalable network but which are nonetheless significant). An important advance in this regard is the use of a pair of ensembles for entanglement generation to achieve *a posteriori* teleportation of light to an atomic memory[73].

There has also been considerable effort devoted to the detailed characterization of decoherence for stored atomic excitation and entanglement[50,65,73]. Decoherence of entanglement between distinct atomic ensembles has been observed in the decay of the violation of Bell's inequality[65] and of the fidelity for teleportation[73]. By measuring concurrence $C(t)$, quantitative characterizations of the relationship between the global evolution of the entangled state and the temporal dynamics of various local correlations were also able to be made[50].

Extending entanglement for quantum networks

The entangled states that have been created so far both in cavity QED and by using the DLCZ protocol are between pairs of systems (known as bipartite entanglement) for which there are definitive procedures for operational verification[72]. The creation of more-general classes of entangled state shared between more than two nodes would be of great interest. However, as researchers progress towards more-complex quantum networks, the issue of entanglement verification becomes increasingly problematic. At present, the theoretical tools and experimental

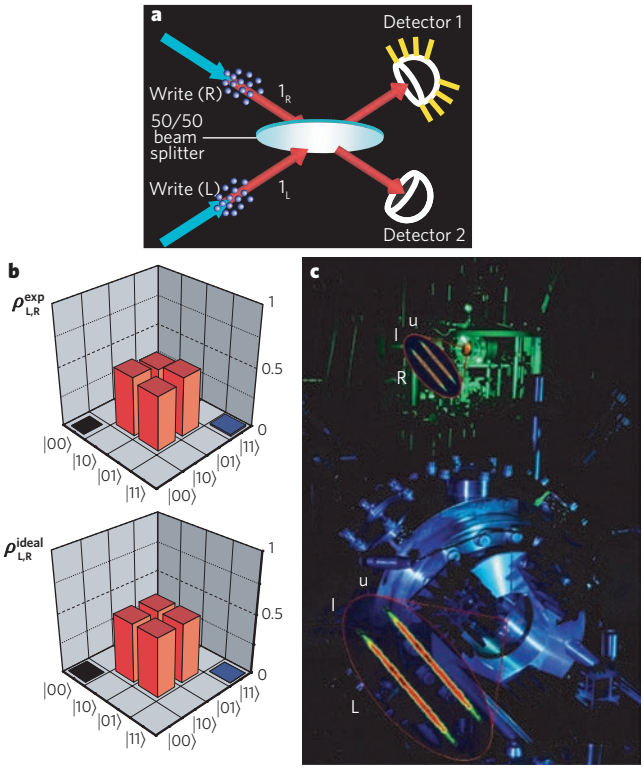


Figure 3 | Fundamentals of the DLCZ protocol. A realistic scheme for entanglement distribution by way of a quantum-repeater architecture was proposed by Duan, Lukin, Cirac and Zoller and is known as the DLCZ protocol[13]. **a**, Measurement-induced entanglement between two atomic ensembles[13,49], L and R, is shown. Synchronized laser pulses incident on the ensembles (denoted write beams, blue arrows) generate small amplitudes for optical fields from spontaneous Raman scattering[48]; these fields are denoted 1_L and 1_R (red arrows). These fields interfere at a 50/50 beam splitter, with outputs directed to two single-photon detectors. A measurement event at either detector (shown for detector 1) projects the ensembles into the entangled state $|\Psi_{L,R}\rangle$ with one quantum of excitation shared remotely between the ensembles. Entanglement is stored in the quantum memory provided by the ensembles and can subsequently be converted to propagating light pulses by a set of 'read' laser pulses (Box 3). **b**, Experimentally determined components of the density matrix $\rho^{exp}_{L,R}$ for entanglement between two atomic ensembles are shown[50], corresponding to concurrence $C = 0.9 \pm 0.3$, where $C = 0$ for an unentangled state. The first number in each ket refers to the excitation number for the ensemble L, and the second is for the ensemble R. For comparison, the density matrix $\rho^{ideal}_{L,R}$ for the ideal state $|\Psi_{L,R}\rangle$ is shown, with concurrence $C = 1$. **c**, The laboratory set-up is shown for the entanglement of two pairs of atomic ensembles to generate the functional quantum nodes L and R, which are separated by 3 m (ref. 65). Each of the four elongated ovals shows a cylinder of 10^5 caesium atoms, which forms an atomic ensemble at each site. Entangled states between the upper u and lower l pairs at the L and R nodes, $|\Psi^u_{L,R}\rangle|\Psi^l_{L,R}\rangle$, are generated and stored in an asynchronous manner for each pair (u and l) as is the case in panel **a**. Atomic excitations for the pairs L_u, L_l and R_u, R_l are subsequently converted to flying photons at each node, with a polarization encoding that results in violation of Bell's inequality[65]. The entire experiment functions under the quantum control of single photon detection events.

capabilities for characterizing the general states of quantum networks do not exist.

Perhaps surprisingly, a non-trivial task will be to find out whether a quantum network 'works'. As moderately complex quantum networks are realized in the laboratory, it will become increasingly more difficult to assess the characteristics of a network quantitatively, including whether entanglement extends across the whole network. One strategy, motivated by the underlying physical processes of the network, could be to try to determine the density matrix $\rho(t)$ for the network. However, this approach would fail because of the exponential growth in $\rho(t)$ with the size of a network.

An alternative strategy could be based on more functional issues of algorithmic capability. An attempt could be made to implement a quantum algorithm for computation or communication to test whether the purported quantum network has greater capabilities than any classical counterpart. This course is, however, problematic because the advantage of a quantum network might only be realized above some threshold in the size of the network. Furthermore, from an experimental perspective, this strategy does not offer much in the way of diagnostics for 'fixing' the network when it fails.

Another, less obvious, approach might be to adopt more seriously the perspective of a quantum network as a quantum many-body system and to search for more 'physical' characteristics of the network (for example, the scaling behaviour of pair correlation functions and multipartite entanglement). Indeed, an active area of research is the nature of entanglement for systems that undergo quantum phase transitions, and there have been pioneering advances in the study of one-dimensional spin chains[74].

Conclusion

Progress has been made towards the development of quantum networks, but the current state of the art is primitive relative to that required for the robust and scalable implementation of sophisticated network protocols, whether over short or long distances. The realization of quantum memories, local quantum processing, quantum repeaters and error-corrected teleportation are ambitious goals. Nevertheless, there is considerable activity directed towards these goals worldwide.

Here cavity-QED-based networks and networks implemented using the DLCZ protocol were considered separately, but it is clear that quantum networks will evolve as heterogeneous entities. For example, the same protocol that creates the entanglement between the two ensembles shown in Fig. 3a can be used to create an entangled state with one excitation shared between an atom in a cavity and an atomic ensemble. A crucial task will be the development of unambiguous procedures for verifying entanglement, a non-trivial undertaking that has not always been carried out correctly[72].

I have used quantum networks as a unifying theme, but the research described here has broader value, including advancing the understanding of quantum dynamical systems and, for the cases considered here, creating new physics from controlled nonlinear interactions of single photons and atoms. These are exciting times in quantum information science as researchers pass from the regime of individual building blocks (for example, a single atom–cavity system) to the realm of complex quantum systems that are assembled block by block from many such units. ■

1. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, UK, 2000).
2. Zoller, P. *et al.* Quantum information processing and communication. Strategic report on current status, visions and goals for research in Europe. *Eur. Phys. J. D* **36,** 203–228 (2005).
3. Bennett, C. H., Brassard, G. & Ekert, A. K. Quantum cryptography. *Sci. Am.* **267** (4), 50–57 (1992).
4. Bouwmeester, D., Ekert, A. & Zeilinger, A. (eds) *The Physics of Quantum Information* (Springer, Berlin, 2000).
5. Giovannetti, V., Lloyd, S. & Maccone, L. Quantum-enhanced measurements: beating the standard quantum limit. *Science* **306,** 1330–1336 (2004).
6. Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78,** 3221–3224 (1997).
7. Preskill, J. P. Plug-in quantum software. *Nature* **402,** 357–358 (1999).
8. Copsey, D. *et al.* Toward a scalable, silicon-based quantum computing architecture. *IEEE J. Quantum Electron.* **9,** 1552–1569 (2003).
9. Illuminati, D. Light does matter. *Nature Phys.* **2,** 803–804 (2006).
10. Duan, L.-M., Wang, B. & Kimble, H. J. Robust quantum gates on neutral atoms with cavity-assisted photon scattering. *Phys. Rev. A* **72,** 032333 (2005).
11. Acín, A., Cirac, J. I. & Lewenstein, M., Entanglement percolation in quantum networks. *Nature Phys.* **3,** 256–259 (2007).
12. Miller, R. *et al.* Trapped atoms in cavity QED: coupling quantized light and matter. *J. Phys. B.* **38,** S551–S565 (2005).
13. Duan, L.-M., Lukin, M. D., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414,** 413–418 (2001).
14. Parkins, A. S., Marte, P., Zoller, P. & Kimble, H. J. Synthesis of arbitrary quantum states via adiabatic transfer of Zeeman coherence. *Phys. Rev. Lett.* **71,** 3095–3098 (1993).

15. van Enk, S. J., Cirac, J. I. & Zoller, P. Photonic channels for quantum communication. *Science* **279,** 205–208 (1998).
16. Lukin, M. D. Trapping and manipulating photon states in atomic ensembles. *Rev. Mod. Phys.* **75,** 457–472 (2003).
17. Briegel, H.-J., Dür, W., Cirac, J. I. & Zoller, P. Quantum repeaters: the role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81,** 5932–5935 (1998).
18. Boozer, A. D., Boca, A., Miller, R., Northup, T. E. & Kimble, H. J. Reversible state transfer between light and a single trapped atom. *Phys. Rev. Lett.* **98,** 193601 (2007).
19. Choi, K. S., Deng, H., Laurat, J. & Kimble, H. J. Mapping photonic entanglement into and out of a quantum memory. *Nature* **452,** 67–71 (2008).
20. Berman, P. (ed.) *Cavity Quantum Electrodynamics* (Academic, San Diego, 1994).
21. Wilk, T., Webster, S. C., Kuhn, A. & Rempe, G. Single-atom single-photon quantum interface. *Science* **317,** 488–490 (2007).
22. Walther, H. Quantum optics of single atoms. *Fortschr. Phys.* **52,** 1154–1164 (2004).
23. Meschede, D., Walther, H. & Mueller, G. One-atom maser. *Phys. Rev. Lett.* **54,** 551–554 (1985).
24. Raimond, J. M. *et al.* Probing a quantum field in a photon box. *J. Phys. B* **38,** S535–S550 (2005).
25. Guerlin, C. *et al.* Progressive field-state collapse and quantum non-demolition photon counting. *Nature* **448,** 889–893 (2007).
26. Vahala, K. J. Optical microcavities. *Nature* **424,** 839–846 (2004).
27. Khitrova, G., Gibbs, H. M., Kira, M., Koch, S. W. & Scherer, A. Vacuum Rabi splitting in semiconductors. *Nature Phys.* **2,** 81–90 (2006).
28. Schoelkopf, R. J. & Girvin, S. M. Wiring up quantum systems. *Nature* **451,** 664–669 (2008).
29. Ye, J., Vernooy, D. W. & Kimble, H. J. Trapping of single atoms in cavity QED. *Phys. Rev. Lett.* **83,** 4987–4990 (1999).
30. Hijlkema, M. *et al.* A single-photon server with just one atom. *Nature Phys.* **3,** 253–255 (2007).
31. Fortier, K. M., Kim, S. Y., Gibbons, M. J. Ahmadi, P. & Chapman, M. S. Deterministic loading of individual atoms to a high-finesse optical cavity. *Phys. Rev. Lett.* **98,** 233601 (2007).
32. Ye, J., Kimble, H. J. & Katori, H. Quantum state engineering and precision metrology using state-insensitive light traps. *Science* (in the press).
33. Boozer, A. D., Boca, A., Miller, R., Northup, T. E. & Kimble, H. J. Cooling to the ground state of axial motion for one atom strongly coupled to an optical cavity. *Phys. Rev. Lett.* **97,** 083602 (2006).
34. Keller, M., Lange, B., Hayasaka, K., Lange, W. & Walther, H. Continuous generation of single photons with controlled waveform in an ion-trap cavity system. *Nature* **431,** 1075–1078 (2004).
35. Duan, L.-M. & Kimble, H. J. Scalable photonic quantum computation through cavity-assisted interactions. *Phys. Rev. Lett.* **92,** 127902 (2004).
36. McKeever, J. *et al.* Deterministic generation of single photons from one atom trapped in a cavity. *Science* **303,** 1992–1994 (2004).
37. Bergmann, K., Theuer, H. & Shore, B. W. Coherent population transfer among quantum states of atoms and molecules. *Rev. Mod. Phys.* **70,** 1003–1025 (1998).
38. Lounis, B. & Orrit, M. Single-photon sources. *Rep. Prog. Phys.* **68,** 1129–1179 (2005).
39. Lange, W. & Kimble, H. J. Dynamic generation of maximally entangled photon multiplets by adiabatic passage. *Phys. Rev. A* **61,** 063817 (2000).
40. Duan, L.-M. & Kimble, H. J. Efficient engineering of multiatom entanglement through single-photon detections. *Phys. Rev. Lett.* **90,** 253601 (2003).
41. Sun, B., Chapman, M. S. & You, L. Atom–photon entanglement generation and distribution. *Phys. Rev. A* **69,** 042316 (2004).
42. Englund, D., Faraon, A., Zhang, B., Yamamoto, Y. & Vuckovic, J. Generation and transfer of single photons on a photonic crystal chip. *Opt. Express* **15,** 5550–5558 (2007).
43. Schön, C., Solano, E., Verstraete, F., Cirac, J. I. & Wolf, M. M. Sequential generation of entangled multiqubit states. *Phys. Rev. Lett.* **95,** 110503 (2005).
44. Parkins, A. S. & Kimble, H. J. Quantum state transfer between motion and light. *J. Opt. B* **1,** 496–504 (1999).
45. Julsgaard, B., Kozhekin, A. & Polzik, E. S. Experimental long-lived entanglement of two macroscopic objects. *Nature* **413,** 400–403 (2001).
46. Sherson, J. F. *et al.* Quantum teleportation between light and matter. *Nature* **443,** 557–560 (2006).
47. Cerf, N. J., Leuchs, G. & Polzik, E. S. (eds) *Quantum Information with Continuous Variables of Atoms and Light* (World Scientific, New Jersey, 2007).
48. Raymer, M. G., Walmsley, I. A., Mostowski, J. & Sobolewska, B. Quantum theory of spatial and temporal coherence properties of stimulated Raman scattering. *Phys. Rev. A* **32,** 332–344 (1985).
49. Chou, C.-W. *et al.* Measurement-induced entanglement for excitation stored in remote atomic ensembles. *Nature* **438,** 828–832 (2005).
50. Laurat, J., Choi, K. S., Deng, H., Chou, C.-W. & Kimble, H. J. Heralded entanglement between atomic ensembles: preparation, decoherence, and scaling. *Phys. Rev. Lett.* **99,** 180504 (2007).
51. Laurat, J. *et al.* Towards experimental entanglement connection with atomic ensembles in the single excitation regime. *New J. Phys.* **9,** 207–220 (2007).
52. Jiang, L., Taylor, J. M. & Lukin, M. D. Fast and robust approach to long-distance quantum communication with atomic ensembles. *Phys. Rev. A* **76,** 012301 (2007).
53. Sangouard, N. *et al.* Robust and efficient quantum repeaters with atomic ensembles and linear optics. Preprint at http://arxiv.org/abs/0802.1475) (2008).
54. Kuzmich, A. *et al.* Generation of nonclassical photon pairs for scalable quantum communication with atomic ensembles. *Nature* **423,** 731–734 (2003).
55. Balic, V., Braje, D. A., Kolchin, P., Yin, G. Y. & Harris, S. E. Generation of paired photons with controllable waveforms. *Phys. Rev. Lett.* **94,** 183601 (2005).
56. van der Wal, C. H. *et al.* Atomic memory for correlated photon states. *Science* **301,** 196–200 (2003).
57. Chou, C. W., Polyakov, S. V., Kuzmich, A. & Kimble, H. J. Single-photon generation from stored excitation in an atomic ensemble. *Phys. Rev. Lett.* **92,** 213601 (2004).
58. Laurat, J. *et al.* Efficient retrieval of a single excitation stored in an atomic ensemble. *Opt. Express* **14,** 6912–6918 (2006).
59. Thompson, J. K., Simon, J., Loh, H. & Vuletic, V. A high-brightness source of narrowband, identical-photon pairs. *Science* **313,** 74–77 (2006).
60. Matsukevich, D. N. *et al.* Deterministic single photons via conditional quantum evolution. *Phys. Rev. Lett.* **97,** 013601 (2006).
61. Chen, S. *et al.* Deterministic and storable single-photon source based on a quantum memory. *Phys. Rev. Lett.* **97,** 173004 (2006).
62. Simon, J., Tanji, H., Thompson, J. K. & Vuletic, V. Interfacing collective atomic excitations and single photons. *Phys.Rev. Lett.* **98,** 183601 (2007).
63. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80,** 2245–2248 (1998).
64. Felinto, D. *et al.* Conditional control of the quantum states of remote atomic memories for quantum networking. *Nature Phys.* **2,** 844–848 (2006).
65. Chou, C.-W. *et al.* Functional quantum nodes for entanglement distribution over scalable quantum networks. *Science* **316,** 1316–1320 (2007).
66. Harris, S. E. Electromagnetically induced transparency. *Phys. Today* **50,** 36–40 (1997).
67. Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* **409,** 490–493 (2001).
68. Phillips, D. F., Fleischhauer, A., Mair, A., Walsworth, R. L. & Lukin, M. D. Storage of light in atomic vapor. *Phys. Rev. Lett.* **86,** 783–786 (2001).
69. Chanelière, T. *et al.* Storage and retrieval of single photons transmitted between remote quantum memories. *Nature* **438,** 833–836 (2005).
70. Eisaman, M. D. *et al.* Electromagnetically induced transparency with tunable single-photon pulses. *Nature* **438,** 837–841 (2005).
71. Simon, J., Tanji, H., Ghosh, S. & Vuletic, V. Single-photon bus connecting spin-wave quantum memories. *Nature Phys.* **3,** 765–769 (2007).
72. van Enk, S. J., Lütkenhaus, N. & Kimble, H. J. Experimental procedures for entanglement verification. *Phys. Rev. A* **75,** 052318 (2007).
73. Chen, Y.-A. *et al.* Memory-built-in quantum teleportation with photonic and atomic qubits. *Nature Phys.* **4,** 103–107 (2008).
74. Vidal, G., Latorre, J. I., Rico, E. & Kitaev, A. Entanglement in quantum critical phenomena. *Phys. Rev. Lett.* **90,** 227902 (2003).
75. Aoki, T. *et al.* Observation of strong coupling between one atom and a monolithic microresonator. *Nature* **443,** 671–674 (2006).
76. Dayan, B. *et al.* A photon turnstile dynamically regulated by one atom. *Science* **319,** 1062–1065 (2008).
77. Spillane, S. M. *et al.* Ultrahigh-*Q* toroidal microresonators for cavity quantum electrodynamics. *Phys. Rev. A* **71,** 013817 (2005).
78. Trupke, M. *et al.* Atom detection and photon production in a scalable, open, optical microcavity. *Phys. Rev. Lett.* **99,** 063601 (2007).
79. Park, Y.-S., Cook, A. K. & Wang, H. Cavity QED with diamond nanocrystals and silica microspheres. *Nano Lett.* **6,** 2075–2079 (2006).
80. Armani, D. K., Kippenberg, T. J., Spillane, S. M. & Vahala, K. J. Ultra-high-*Q* toroid microcavity on a chip. *Nature* **421,** 925–928 (2003).
81. Spillane, S. M., Kippenberg, T. J., Painter, O. J. & Vahala, K. J. Ideality in a fiber-taper-coupled microresonator system for application to cavity quantum electrodynamics. *Phys. Rev. Lett.* **91,** 043902 (2003).
82. Braginsky, V. B., Gorodetsky, M. L. & Ilchenko, V. S. Quality-factor and nonlinear properties of optical whispering-gallery modes. *Phys. Lett. A* **137,** 393–397 (1989).

Acknowledgements I am grateful for the contributions of members of the Caltech Quantum Optics Group, especially K. S. Choi, B. Dayan and R. Miller. I am indebted to J. P. Preskill and S. J. van Enk for critical insights. My research is supported by the National Science Foundation, IARPA and Northrop Grumman Space Technology.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (hjkimble@caltech.edu).

Superconducting quantum bits

John Clarke[1,2] & Frank K. Wilhelm[3]

Superconducting circuits are macroscopic in size but have generic quantum properties such as quantized energy levels, superposition of states, and entanglement, all of which are more commonly associated with atoms. Superconducting quantum bits (qubits) form the key component of these circuits. Their quantum state is manipulated by using electromagnetic pulses to control the magnetic flux, the electric charge or the phase difference across a Josephson junction (a device with nonlinear inductance and no energy dissipation). As such, superconducting qubits are not only of considerable fundamental interest but also might ultimately form the primitive building blocks of quantum computers.

The theory of quantum mechanics was originally developed to account for the observed behaviour of electrons in atoms. More than 80 years later, it is being used to explain the behaviour of superconducting circuits that can be hundreds of nanometres wide and can contain trillions of electrons. The quantum nature of these circuits is observable because they can be engineered to be isolated from the electrical environment and are thus represented by a single degree of freedom. Significant coupling to other degrees of freedom causes rapid decoherence, destroying the quantum state of the circuit so that it behaves classically. Unlike atoms, these circuits can be designed and constructed to tailor their characteristic frequencies, as well as other parameters. These frequencies can be controlled by adjusting an external parameter, and the coupling energy between two quantum bits (qubits) can be turned on and off at will.

Superconducting quantum circuits are the subject of intense research at present, in part because they have opened up a new area of fundamental science and in part because of their long-term potential for quantum computing. In this review, we begin with a brief discussion of superconductivity and two of the superconducting properties that underlie how qubits operate: flux quantization and Josephson tunnelling. The three fundamental types of superconducting qubit — flux, charge and phase — are then described. This is followed by a review of the real-time, quantum-coherent dynamics of qubits and the limitations imposed by relaxation and decoherence, as well as the mechanisms of decoherence. We then discuss schemes for controlling the coupling between two qubits, a feature that greatly simplifies the implementation of proposed quantum-computing architectures. And we finish by discussing quantum optics on a chip, a new research direction in which the electromagnetic fields associated with control and read-out signals are treated quantum mechanically.

Flux quantization and Josephson tunnelling

Why do superconductors enable atomic-scale phenomena to be observed at the macroscopic level? The reason, as explained elegantly by the theory of Bardeen, Cooper and Schrieffer[1], is that in a given superconductor all of the Cooper pairs of electrons (which have charge $2e$, mass $2m_e$ and spin zero, and are responsible for carrying a supercurrent) are condensed into a single macroscopic state described by a wavefunction $\Psi(\mathbf{r}, t)$ (where $\mathbf{r}$ is the spatial variable and t is time.) Like all quantum-mechanical wavefunctions, $\Psi(\mathbf{r}, t)$ can be written as $|\Psi(\mathbf{r}, t)| \exp[i\phi(\mathbf{r}, t)]$ (where $i = \sqrt{-1}$): that is, as the product of an amplitude and a factor involving the phase ϕ. Furthermore, in 'conventional' superconductors such as Nb, Pb and Al, the quasiparticles (electron-like and hole-like excitations) are separated in energy from the condensate[2] by an energy gap $\Delta_s(T) = 1.76k_BT_c$ (where k_B is the Boltzmann constant and T_c is the superconducting transition temperature). Thus, at temperatures $T << T_c$, the density of quasiparticles becomes exponentially small, as does the intrinsic dissipation for frequencies of less than $2\Delta_s(0)/h$ (where h is Planck's constant) — roughly 10^{11} Hz for Al.

The macroscopic wavefunction leads to two phenomena that are essential for qubits. The first phenomenon is flux quantization. When a closed ring is cooled through its superconducting transition temperature in a magnetic field and the field is then switched off, the magnetic flux Φ in the ring — maintained by a circulating supercurrent — is quantized[2] in integer values of the flux quantum $\Phi_0 \equiv h/2e \approx 2.07 \times 10^{-15}$ T m^2. This quantization arises from the requirement that $\Psi(\mathbf{r}, t)$ be single valued. The second phenomenon is Josephson tunnelling[2]. A Josephson junction consists of two superconductors separated by an insulating barrier of appropriate thickness, typically 2–3 nm, through which Cooper pairs can tunnel coherently. Brian Josephson showed that the supercurrent I through the barrier is related to the gauge-invariant phase difference $\delta(t)$ between the phases of the two superconductors by the current–phase relationship

$$I = I_0 \sin\delta \tag{1}$$

Here I_0 is the maximum supercurrent that the junction can sustain (that is, the critical current). This phase difference is an electrodynamic variable that, in the presence of a potential difference V between the superconductors, evolves in time as

$$\hbar\dot{\delta} = \hbar\omega = 2eV \tag{2}$$

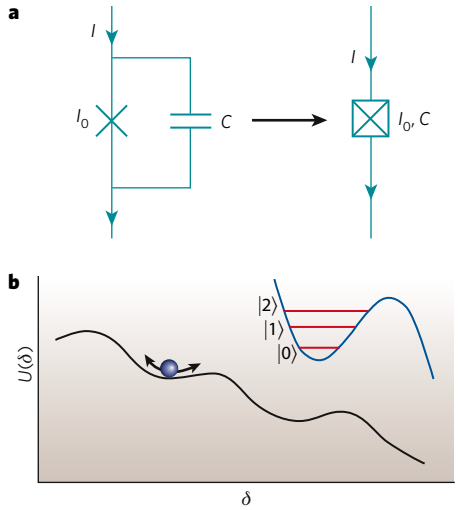
where $\hbar = h/2\pi$ and ω is the angular frequency at which the supercurrent oscillates. The dynamical behaviour of Josephson junctions is described in Box 1.

The variables have, so far, been regarded as being classical, but to show quantum-mechanical behaviour, these variables must be replaced by operators. The two relevant operators are that for δ, which is associated with the Josephson coupling energy $E_j \equiv I_0\Phi_0/2\pi$, and that for the Cooper-pair number difference N across the capacitance, which is associated with the charging energy $E_c \equiv (2e)^2/2C$, where C is the junction capacitance.

Furthermore — just like the familiar position and momentum operators x and p_x — the operators for δ and for the charge on the capacitor Q

[1]Department of Physics,366 LeConte Hall, University of California, Berkeley, California 94720, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA. [3]Institute for Quantum Computing, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada.

Box 1 | The Josephson junction as a nonlinear circuit element



Equations (1) and (2) contain the crucial information that the Josephson junction is a dissipationless device with a nonlinear inductance. It is these unique features that make the junction the primitive building block of all superconducting qubits.

The nonlinear inductance is easily deduced by noting that the time derivative of equation (1) yields $\dot{I} = (I_0\cos\delta)\dot{\delta} = (I_0\cos\delta)\omega = V(2eI_0/\hbar)\cos\delta$ from equation (2). Invoking Faraday's law $V = -L\dot{I}$ (where L is the inductance) then leads to the Josephson inductance

$$|L_j| = \Phi_0/(2\pi I_0\cos\delta) = \Phi_0/2\pi(I_0^2 - I^2)^{1/2} \text{ (where } I < I_0) \qquad (6)$$

The Josephson junction, denoted by an X in panel **a** of the figure, has an intrinsic capacitance C; this combination is often denoted by an X in a box. I_0 denotes the critical current. It is immediately apparent from equation (6) that the junction is also a nonlinear oscillator with a resonant angular frequency $\omega_p(I) = (L_jC)^{-1/2} = (2\pi I_0/\Phi_0C)^{1/2}/(1 - I^2/I_0^2)^{1/4}$.

Considerable insight into the dynamics of a Josephson junction can be gleaned by considering the flow of a current J through the junction: $J = I_0\sin\delta + C\dot{V}$. Writing $\dot{V} = (\hbar/2e)\ddot{\delta}$ and rearranging this yields $(\hbar C/2e)\ddot{\delta} = I - I_0\sin\delta = -(2e/\hbar)\partial U/\partial\delta$. $U \equiv -(\Phi_0/2\pi)\partial U/\partial\delta$ is the potential of a tilted washboard for a particle of mass $\hbar C/2e$ (as illustrated in panel **b** of the figure). In the absence of fluctuations, for $I < I_0$ the particle remains trapped in one of the potential wells; classically, it oscillates in the well at the plasma oscillation frequency $\omega_p(I)/2\pi$. Thus, $\langle\dot{\delta}\rangle = 0$, and the junction is in the zero-voltage state; in the quantum picture, the energy in the well is quantized, as shown in the inset (figure, panel **b**). By contrast, when I is increased so that $I > I_0$, the particle runs down the washboard, $\langle\dot{\delta}\rangle > 0$, and there is a voltage across the junction. When I is subsequently reduced so that $I < I_0$, the particle will continue to propagate until I is close to 0. Thus, the current–voltage characteristic is hysteretic.

are canonically conjugate, as expressed by the commutator bracket $[\delta, Q] = i2e$. The fact that δ and Q are subject to Heisenberg's uncertainty principle has far-reaching consequences. On the one hand, when $E_j >> E_c$, δ is well defined, and Q has large quantum fluctuations; therefore, the Josephson behaviour of the junction dominates. On the other hand, when $E_j << E_c$, N is well defined, and δ has large quantum fluctuations; therefore, the charging behaviour of the capacitor dominates. Using these ideas, the parameters of superconducting quantum circuits can be designed[3].

The first evidence of quantum behaviour in a Josephson junction came from experiments in which macroscopic quantum tunnelling was found to occur and energy levels were shown to be quantized. In macroscopic quantum tunnelling[4,5], the junction tunnels from the ground state $|0\rangle$ (Box 1 figure), when $I < I_0$, through the potential barrier that separates it from its neighbouring energy well, which is at a lower energy. Then, the particle runs freely down the washboard potential, generating a voltage $2\Delta_s/e$ that is readily detected. These results[5] were found to be in strong agreement with theory[6]. Energy quantization[7] was found in the initial well by irradiating the junction with microwaves. The escape rate from the zero-voltage state was increased when the microwave frequency f_m corresponded to the energy difference between two adjacent energy levels. A crucial point is that the anharmonic nature of the well, which results from the nonlinear inductance of Josephson junctions (equation (6), Box 1), causes the energy spacing to decrease as the quantum number progressively increases, so each transition has a distinct frequency. If the well were harmonic, the energy spacings would be identical, and the quantum case would not be distinguishable from the classical case.

These experiments showed unequivocally that δ is a quantum variable. The next step in the demonstration of macroscopic quantum physics was to implement devices showing the superposition of two quantum states $|\Psi_1\rangle$ and $|\Psi_2\rangle$ in the form $|\Psi\rangle = \alpha|\Psi_1\rangle + \beta|\Psi_2\rangle$, as first proposed by Anthony Leggett[8] in the 1980s in his discussion of macroscopic quantum coherence in superconducting devices. In 1997, Yasunobu Nakamura *et al.*[9] carried out the first such experiment on a charge qubit, showing spectroscopically the superposition of the Cooper-pair states $|n\rangle$ and $|n+1\rangle$, where the integer n is the quantum number specifying the number of Cooper pairs. Subsequently, in 2000, Jonathan Friedman *et al.*[10] and Caspar van der Wal *et al.*[11] showed the superposition of states in a flux qubit. A flux qubit consists of a superconducting loop interrupted by one[10] or three[11] Josephson junctions. The two quantum states are flux pointing up and flux pointing down or, equivalently, supercurrent flowing in an anticlockwise direction and supercurrent flowing in a clockwise direction. In 2002, Denis Vion *et al.*[12] described 'quantronium', a qubit in which two small junctions are connected by a superconducting island, involving the superposition of the Cooper-pair states $|n\rangle$ and $|n+1\rangle$. Also in 2002, John Martinis *et al.*[13] demonstrated a phase qubit, a reinvention of the device used earlier to observe quantized energy levels[7]. The relevant quantum states are the ground state and the first excited state. Some of the experimental difficulties encountered when operating superconducting qubits are described in Box 2.

Flux qubits

A flux qubit, as indicated earlier, consists of a superconducting loop interrupted by one[10] or three[11] Josephson junctions (Fig. 1a). Although both designs function similarly, we focus on the three-junction design, which has been adopted more widely. In this device, one junction is smaller in area and thus has a smaller critical current than the other two, which function to increase the inductance of the loop. The small junction has a large value for E_j/E_c, typically 50, so the phase difference δ (or, equivalently, the magnetic flux Φ in the loop) is the relevant quantum variable. The two quantum states are magnetic flux pointing up $|\uparrow\rangle$ and magnetic flux pointing down $|\downarrow\rangle$ or, equivalently, anticlockwise qubit supercurrent I_q circulating in the loop and clockwise supercurrent. The qubit is represented by a double-well potential, which is generally asymmetrical. The two states are coupled by the quantum-mechanical tunnelling of δ through the barrier separating the wells, giving rise to the superposition of the two basis states

$$|\Psi\rangle = \alpha|\uparrow\rangle \pm \beta|\downarrow\rangle \qquad (3)$$

When the externally applied magnetic flux $\Phi_e = \Phi_0/2$, the double-well potential becomes symmetrical (Fig. 1b), and the two eigenfunctions become symmetrical and antisymmetrical superpositions of the two basis states, with $\alpha = \beta = 1/\sqrt{2}$. At this degeneracy point, the splitting of the energy levels of the ground state $|0\rangle$ and the first excited state $|1\rangle$ is Δ; away from the degeneracy point, the energy difference is

$$\nu = (\Delta^2 + \varepsilon^2)^{1/2} \qquad (4)$$

where $\varepsilon = 2I_q(\Phi_e - \Phi_0/2)$ (Fig. 1c). The probabilities of observing the states $|\uparrow\rangle$ and $|\downarrow\rangle$ in the ground and first excited states as a function

Box 2 | Experimental issues with superconducting qubits

Experiments on superconducting qubits are challenging. Most superconducting qubits are created by using electron-beam lithography, need millikelvin temperatures and an ultralow-noise environment to operate, and can be studied only by using very sensitive measurement techniques.

Superconducting qubits generally require Josephson junctions with dimensions of the order of $0.1 \times 0.1\,\mu m^2$ — corresponding to a self-capacitance of about 1 fF — and are patterned by using shadow evaporation and electron-beam lithography[79]; an exception is the phase qubit, which typically has a junction of $1 \times 1\,\mu m^2$ and can be patterned photolithographically. The Josephson junctions are usually Al–Al_xO_y–Al (where $x \leq 2$ and $y \leq 3$), and the oxidation must be controlled to yield relatively precise values of E_J and E_C. Because qubit frequencies are usually 5–10 GHz (which corresponds to 0.25–0.5 K), the circuits are operated in dilution refrigerators, typically at temperatures of 10–30 mK, to minimize thermal population of the upper state.

Great efforts are made to attenuate external electrical and magnetic noise. The experiment is invariably enclosed in a Faraday cage — either a shielded room or the metal Dewar of the refrigerator with a contiguous metal box on top. The electrical leads that are connected to the qubits and their read-out devices are heavily filtered or attenuated. For example, lines carrying quasistatic bias currents usually have multiple low-pass filters at the various temperature stages of the refrigerator. These include both inductor–capacitor and resistor–capacitor filters that operate up to a few hundred megahertz, as well as wires running through copper powder, which results in substantial loss at higher frequencies[5]. The overall attenuation is typically 200 dB. Finally, the read-out process for probing a quantum system is very delicate.

of Φ_e are shown in Fig. 1d. At the degeneracy point, the probability of observing either state is ½. As Φ_e is reduced, the probability of observing $|\uparrow\rangle$ increases while that of observing $|\downarrow\rangle$ decreases.

The first observation of quantum superposition in a flux qubit was made spectroscopically. The state of the flux qubit is measured with a d.c. superconducting quantum interference device (SQUID)[14]. This device consists of two Josephson junctions, each with critical current I_0, connected in parallel on a superconducting loop of inductance L. The critical current of the SQUID $I_c(\Phi_s)$ is periodic in the externally applied magnetic flux Φ_s with period Φ_0. In the limit $\beta_L \equiv 2LI_0/\Phi_0 << 1$ in which the Josephson inductance dominates the geometrical inductance, the critical current for $\Phi_s = (m + ½)\Phi_0$ (m is an integer) is reduced to almost zero, and the flux dependence of the critical current takes the approximate form[14] $I_c(\Phi_s) \approx 2I_0|\cos(\pi\Phi_s/\Phi_0)|$. Thus, by biasing the SQUID with a constant magnetic flux near $\Phi_0/2$, and measuring the critical current, the changes in flux produced by a nearby qubit can be measured with high sensitivity. In most experiments with qubits, a pulse of current is applied to the SQUID, which either remains in the zero-voltage state or makes a transition to the voltage state, producing a voltage $2\Delta_s/e$. Because its current–voltage characteristic is hysteretic, the SQUID remains at this voltage until the current bias has been removed, allowing researchers to determine whether the SQUID has switched. For sufficiently small current pulses, the probability of the SQUID switching is zero, whereas the probability is one for sufficiently large pulses. The switching event is a stochastic process and needs to be repeated many times for the flux in the SQUID to be measured accurately.

The first step in spectroscopic observation of quantum superposition is to determine the height of the current pulse at which the SQUID switches — with, for example, a probability of ½ — as a function of Φ_e over a narrow range (perhaps $\pm 5m\Phi_0$). Subsequently, a pulse of microwave flux is applied at frequency f_m, which is of sufficient amplitude and duration to equalize the populations of the ground state and first excited state when the energy-level splitting difference $\nu = hf_m$. Assuming that $|\uparrow\rangle$ is measured, then, on resonance, there will be a peak in the switching probability for $\Phi_e < \Phi_0/2$ and a corresponding dip for $\Phi_e > \Phi_0/2$. An example of these results[11,15] is shown in Fig. 2. The configuration of the qubit and the SQUID is shown in Fig. 2a, and the peaks and dips in the amplitude of the switching current

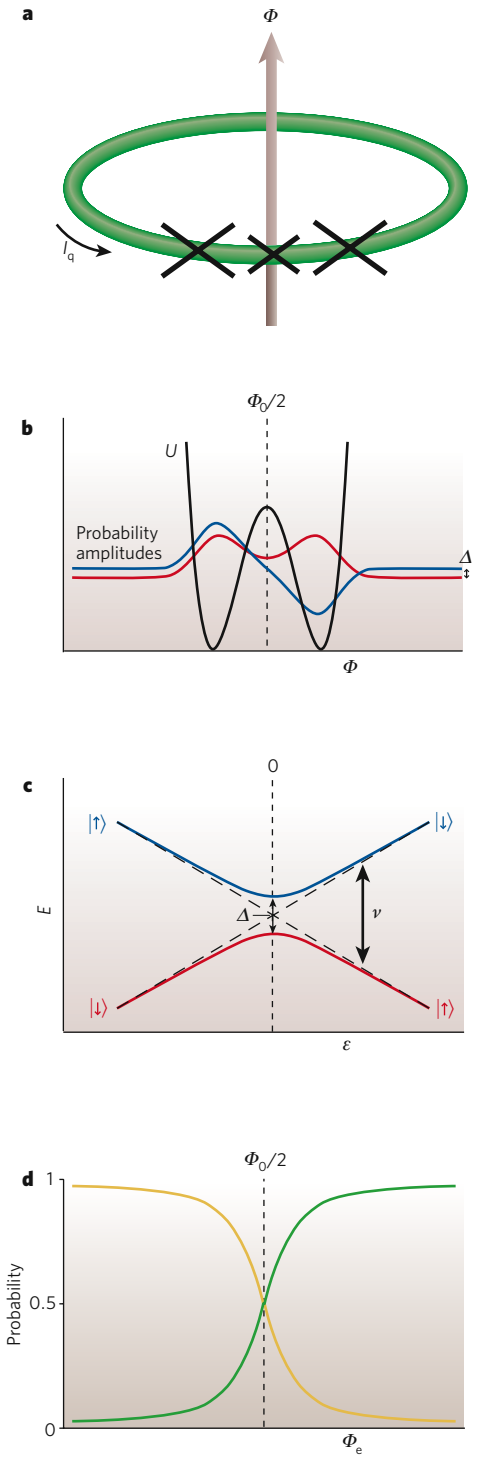


Figure 1 | The theory underlying flux qubits. **a**, Flux qubits consist of a superconducting loop interrupted by either one or three (shown) Josephson junctions. The two quantum states are magnetic flux Φ pointing up $|\uparrow\rangle$ and Φ pointing down $|\downarrow\rangle$ or, equivalently, supercurrent I_q circulating in the loop anticlockwise and I_q circulating clockwise. **b**, The double-well potential (black) versus total flux Φ contained in a flux qubit is shown. The two wells are symmetrical when the externally applied magnetic flux Φ_e is $(n + ½)\Phi_0$, where n is an integer (n = 0 in this case). The coloured curves are the eigenfunctions (probability amplitudes) for the ground state (symmetrical; red) and first excited state (antisymmetrical; blue). **c**, The energy E of the two superpositions states in **b** versus the energy bias $\varepsilon = 2I_q(\Phi_e - \Phi_0/2)$ is shown. The diagonal dashed black lines show the classical energies. The energy-level splitting is Δ at the degeneracy point, $\varepsilon = 0$, and is ν for $\varepsilon \neq 0$. **d**, The probabilities of the qubit flux pointing up (green) or down (yellow) in the ground state versus applied flux are shown.

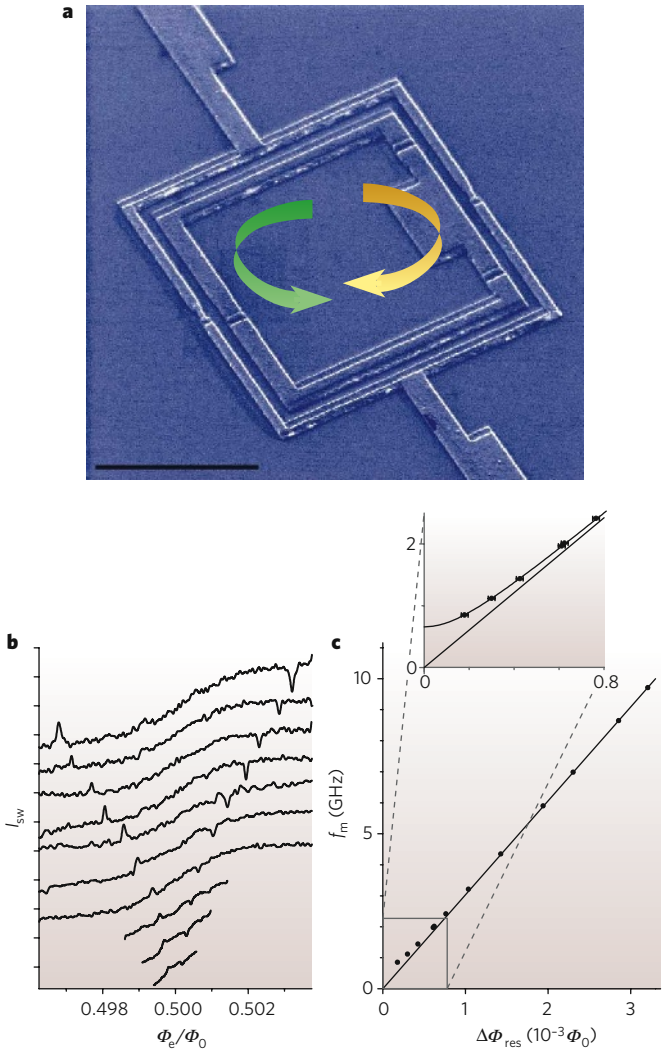


Figure 2 | Experimental properties of flux qubits. a, The configuration of the original three-junction flux qubit is shown. Arrows indicate the current flow in the two qubit states (green denotes $|\uparrow\rangle$, and yellow denotes $|\downarrow\rangle$). Scale bar, 3 μm. (Image courtesy of C. H. van der Wal, Rijksuniversiteit Groningen, the Netherlands). **b**, Radiation of microwave frequency f_m induces resonant peaks and dips in the switching current I_{sw} with respect to the externally applied magnetic flux Φ_e normalized to the flux quantum Φ_0. Frequencies range from 9.711 GHz to 0.850 GHz. Tick marks on the *y* axis show steps of 0.4 nA. (Panel reproduced, with permission, from ref. 15.) **c**, Microwave frequency f_m is plotted against half of the separation in magnetic flux, $\Delta\Phi_{res}$, between the peak and the dip at each frequency. The line is a linear fit through the data at high frequencies and represents the classical energy. The inset is a magnified view of the lower part of the graph; the curved line in the inset is a fit to equation (4). The deviation of the data points from the straight line demonstrates quantum coherence of the $|\uparrow\rangle$ and $|\downarrow\rangle$ flux states. (Panel reproduced, with permission, from ref. 15.)

versus applied flux are shown in Fig. 2b for a succession of microwave frequencies. As expected, the difference in the applied flux at which the peaks and dips appear, $2\Delta\Phi_{res}$, becomes greater as the microwave frequency increases. The microwave frequency versus $\Delta\Phi_{res}$ is shown in Fig. 2c. The data have been fitted to equation (4) with $I_q = (\frac{1}{2})dv/d\Phi_e$ in the limit $v >> \Delta$, using Δ as a fitting parameter. The data reveal the existence of an anticrossing (that is, an avoided crossing) at $\Phi_e = \Phi_0/2$.

Charge qubits

A charge qubit (also known as a Cooper-pair box) is shown in Fig. 3a, b. The key component is a tiny superconducting island that is small enough that the electrostatic charging energy required to place a charge of 2*e* on the island at zero voltage, $(2e)^2/2C_\Sigma$, is much greater than the thermal energy k_BT (where $C_\Sigma = C_g + C_j$ is the total capacitance). For $T = 1\,K$, this requires C_Σ to be much less than 1 fF. The Cooper-pair box is connected to ground by a gate capacitance C_g in series with a potential V_g and by a small Josephson junction with $E_j << E_c$. Given their weak connection to the 'outside world', the number of Cooper pairs on the island is a discrete variable *n*. The qubit states correspond to adjacent Cooper-pair number states $|n\rangle$ and $|n+1\rangle$.

To understand how to control a single Cooper pair, it is useful to first examine the electrostatic problem with an infinite junction resistance ($E_j = 0$). The total electrostatic energy of the circuit is $E_{ch} = (2e^2/C_g)(n - n_g)^2$, where $n_g = C_gV_g/2e$ (representing the gate voltage in terms of the gate charge, namely the polarization charge that the voltage induces on the gate capacitor). Although *n* is an integer, n_g is a continuous variable. E_{ch} versus n_g is shown in Fig. 3c for several values of *n*. It should be noted that the curves for *n* and $n+1$ cross at $n_g = n + \frac{1}{2}$, the charge degeneracy point. At this point, the gate polarization corresponds to half a Cooper pair for both charge basis states.

By restoring the Josephson coupling to a small value, the behaviour close to these crossing points is modified. The Josephson junction allows Cooper pairs to tunnel onto the island one by one. The resultant coupling between neighbouring charge states $|n\rangle$ and $|n+1\rangle$ makes the quantum superposition of charge eigenstates analogous to the superposition of flux states in equation (3) (identifying $|\downarrow\rangle = |n\rangle$ and $|\uparrow\rangle = |n+1\rangle$). The next excited charge state is higher in energy by E_c and can safely be neglected. At the charge degeneracy point, where the Josephson coupling produces an avoided crossing, the symmetrical and antisymmetrical superpositions are split by an energy E_j. By contrast, far from this point, $E_c >> E_j$, and the eigenstates are very close to being charge states. Again, the energy level structure is analogous to that of flux qubits, with Δ replaced with E_j and ε with $E_c \times (n_g - n - \frac{1}{2})$. Similarly, the probabilities of measuring the ground state or excited state depend on the gate voltage rather than the applied flux.

To make the qubit fully tunable, the Josephson junction is usually replaced by a d.c. SQUID with low inductance ($\beta_L << 1$). E_j is then adjusted by applying the appropriate magnetic flux, which is kept constant throughout the subsequent measurements.

The read-out of a charge qubit involves detecting the charge on the island to a much greater accuracy than 2*e*. This is accomplished by using a single-electron transistor (SET), a sensitive electrometer[16]. The SET (Fig. 3d), also based on a tiny island, is connected to two superconducting leads by two Josephson junctions. When the voltages across both junctions are close to the degeneracy point ($n_g = n + \frac{1}{2}$), charges cross the junctions to produce a net current flow through the SET. Thus, the current near the degeneracy points depends strongly on the gate voltage (Fig. 3c). Capacitively coupling the Cooper-pair-box island to the SET island makes a contribution to the SET gate voltage so that the SET current strongly depends on the Cooper-pair-box state. This scheme converts the measurement of charge into a measurement of charge transport through a SET. In fact, for small Josephson junctions, this charge transport is usually dissipative, because the phase coherence is destroyed by environmental fluctuations. Thus, the read-out actually involves measuring the resistance of the SET, which depends on the state of the Cooper-pair box. The preferred read-out device is a radio-frequency SET[17], in which a SET is embedded in a resonant circuit. Thus, the *Q* of the resonant circuit is determined by the resistance of the SET and ultimately by the charge on the Cooper-pair box. A pulse of microwaves slightly detuned from the resonant frequency is applied to the radio-frequency SET, and the phase of the reflected signal enables the state of the qubit to be determined.

Many of the initial studies of superconducting qubits involved charge qubits. That crossing is avoided at the degeneracy point was first shown spectroscopically by studying a charge qubit[9], and charge measurements revealed the continuous, quantum-rounded form of the transition between quantum states[18]. The coherent oscillations that occur with time at this avoided energy-level crossing were also first discovered by studying a charge qubit[19].

Cooper-pair boxes are particularly sensitive to low-frequency noise from electrons moving among defects (see the section 'Decoherence') and can show sudden large jumps in n_g. The development of more advanced charge qubits such as the transmon[20] and quantronium[12] has greatly ameliorated this problem. The transmon is a small Cooper-pair box that is made relatively insensitive to charge by shunting the Josephson junction with a large external capacitor to increase E_c and by increasing the gate capacitor to the same size. Consequently, the energy bands of the type shown in Fig. 3c are almost flat, and the eigenstates are a combination of many Cooper-pair-box charge states. For reasons that will be discussed later (see the section 'Decoherence'), the transmon is thus insensitive to low-frequency charge noise at all operating points. At the same time, the large gate capacitor provides strong coupling to external microwaves even at the level of a single photon, greatly increasing the coupling for circuit quantum electrodynamics (QED) (see the section 'Quantum optics on a chip').

The principle by which quantronium operates is shown in Fig. 4a, and an actual circuit is shown in Fig. 4b. The Cooper-pair box involves two Josephson junctions, with a capacitance C_g connected to the island separating them. The two junctions are connected across a third, larger, junction, with a higher critical current, to form a closed superconducting circuit to which a magnetic flux Φ_e is applied. The key to eliminating the effects of low-frequency charge and flux noise is to maintain the qubit at the double degeneracy point at which the two qubit states are (to first order) insensitive to these noise sources. To achieve insensitivity to charge noise, the qubit is operated at $n_g = ½$, where the energy levels have zero slope and the energy-level splitting is E_j (Fig. 3c). Insensitivity to flux noise is achieved by applying an integer number of half-flux quanta to the loop. The success of this optimum working point has been elegantly shown experimentally[21]. The insensitivity to both flux and charge implies, however, that the two states of the qubit cannot be distinguished at the double degeneracy point. To measure the qubit state, a current pulse that moves the qubit away from the flux degeneracy point is applied to the loop, and this produces a clockwise or anticlockwise current in the loop, depending on the state of the qubit. The direction of the current is determined by the third (read-out) junction: the circulating current either adds to or subtracts from the applied current pulse, so the read-out junction switches out of the zero-voltage state at a slightly lower or slightly higher value of the bias current, respectively. Thus, the state of the qubit can be inferred by measuring the switching currents. With the advent of quantronium, much longer relaxation and decoherence times can be achieved than with a conventional Cooper-pair box.

Although this switching read-out scheme is efficient, it has two major drawbacks. First, the resultant high level of dissipation destroys the quantum state of the qubit, making sequential measurements of the state impossible. Second, the temperature of the read-out junction and substrate increase because of the energy that is deposited while the SQUID is in the voltage state — typically for 1 μs — and the equilibrium is not restored for ~1 ms. This limits the rate at which measurements can be made to ~1 kHz, resulting in long data-acquisition times.

These drawbacks have been overcome by the introduction of the Josephson bifurcation amplifier (JBA)[22], a particularly powerful read-out device in which there is no dissipation because the junction remains in the zero-voltage state (Fig. 4c). The JBA exploits the nonlinearity of the Josephson junction when a capacitor is connected across it, resulting in the formation of a resonant (or tank) circuit. When small-amplitude microwave pulses are applied to the resonant circuit, the amplitude and phase of the reflected signal are detected, with the signal strength boosted by a cryogenic amplifier. From this measurement, the resonant frequency of the tank circuit can be determined, then the inductance of the junction — which depends on the current flowing through it — and, finally, the state of the quantronium. For larger-amplitude microwaves, however, the behaviour of the circuit is strongly nonlinear, with the resonance frequency decreasing as the amplitude increases. In particular, strong driving at frequencies slightly below the plasma frequency leads to a bistability: a weak, off-resonance lower branch during which the particle does not explore the nonlinearity, and a high-amplitude response at which frequency matches the driving frequency (Fig. 4d). The two qubit states can be distinguished by choosing driving frequencies and currents that cause the JBA to switch to one response or the other, depending on the qubit state. This technique is extremely fast and, even though it is based on a switching process, it never drives the junction into the voltage state. Furthermore, the JBA remains in the same state after the measurement has been made.

The JBA has been shown to approach the quantum non-demolition (QND) limit[22]. This limit is reached when the perturbation of the quantum state during the measurement does not go beyond that required by the measurement postulate of quantum mechanics, so repeated measurements of the same eigenstate lead to the same outcome[23]. Reaching the QND limit is highly desirable for quantum computing.

A similar scheme that approaches the QND limit has been implemented for the flux qubit, with the single Josephson junction replaced by a read-out SQUID[24]. Dispersive read-out for a flux qubit has also been achieved by inductively coupling a flux qubit to the inductor of a resonant circuit and then measuring the flux state from the shift in the resonance frequency[25].

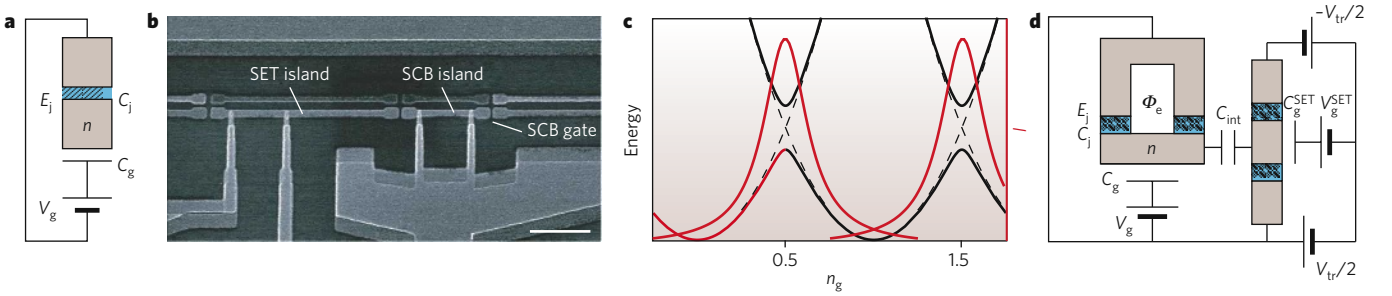


Figure 3 | Charge qubits. a, A single Cooper-pair-box (SCB) circuit is shown. The superconducting island is depicted in brown and the junction in blue. E_j and C_j are the Josephson coupling energy and self-capacitance, respectively, and n is the number of Cooper pairs on the island, which is coupled to a voltage source with voltage V_g by way of a capacitor with capacitance C_g. (Panel reproduced, with permission, from ref. 28.) **b**, A micrograph of a Cooper-pair box coupled to a single-electron transistor (SET) is shown. Scale bar, 1 μm. (Panel reproduced, with permission, from ref. 78.) **c**, Black curves show the energy of the Cooper-pair box as a function of the scaled gate voltage $n_g = C_g V_g/2e$ for different numbers (n) of excess Cooper pairs on the island. The parabola on the far left corresponds to $n = 0$ and the central parabola to $n = 1$. Dashed lines indicate the contribution of the charging energy $E_{ch}(n, n_g)$ alone. The energy-level splitting at $n_g = ½$ is E_j. Red curves show the current I through the SET as a function of n_g. Transport is possible at the charge degeneracy points, where the gate strongly modulates the current. (Panel reproduced, with permission, from ref. 28.) **d**, A charge qubit with two junctions (left) coupled to a SET biased to a transport voltage V_{tr} (right) is shown. The critical current of the junctions coupled to the island is adjusted by means of an externally applied magnetic flux Φ_e. The gate of the SET is coupled to an externally controlled charge induced on the capacitor with capacitance C_g^{SET} by the voltage V_g^{SET}, as well as to the qubit charge by way of the interaction capacitance C_{int}. (Panel reproduced, with permission, from ref. 28.)

Phase qubits

In essence, a phase qubit[13] consists of a single current-biased Josephson junction (Box 1 figure). For a bias current I just below the critical current I_0, the anharmonic potential is approximately cubic, and the energy-level spacing becomes progressively smaller as the quantum number n increases. As I approaches I_0, the (classical) plasma oscillation frequency, $\omega_p(I) = 2^{1/4}(2\pi I_0/\Phi_0 C)^{1/2}(1 - I/I_0)^{1/4}$, decreases slowly, while the potential barrier height, $\Delta U(I) = (2\sqrt{2}\, I_0\Phi_0/3\pi)(1 - I/I_0)^{3/2}$, decreases rapidly. Thus, the probability of escape from the state $|n\rangle$ by macroscopic quantum tunnelling increases exponentially as n increases. The qubit involves transitions between the ground state $|0\rangle$ and the first excited state $|1\rangle$. To measure the quantum state of the qubit, a microwave pulse is applied with frequency $(E_2 - E_1)/h$. If, on the one hand, the qubit is in the state $|1\rangle$, then the pulse excites a transition to the state $|2\rangle$, from which macroscopic quantum tunnelling causes the junction to switch to the voltage state. If, on the other hand, the junction is initially in the state $|0\rangle$, then no such transition occurs. Operation of the phase qubit depends crucially on the anharmonicity of the well potential, which ensures that $E_2 - E_1 < E_1 - E_0$.

The first phase qubit that was designed involved a $10 \times 10\,\mu m^2$ Nb–Al_xO_y–Nb tunnel junction (where $x \leq 2$ and $y \leq 3$), which was created photolithographically. To measure the occupation probability p_1 of the state $|1\rangle$, Martinis *et al.*[13] applied a long microwave pulse of angular frequency $\omega_{10} = (E_1 - E_0)/\hbar$, followed by a read-out pulse of frequency $\omega_{21} = (E_2 - E_1)/\hbar$ (Fig. 5a). If the state $|1\rangle$ is occupied, the second pulse switches the junction to the voltage state, which is detected by a low-noise amplifier. If, conversely, the junction is in the state $|0\rangle$, the probability of switching is very small. As the power P_{10} in the first pulse is increased, the probability of $|1\rangle$ being occupied increases until it reaches a plateau at 0.5. The results of the measurement are shown in Fig. 5a, where p_1 is defined as the ratio of the number of trials in which switching to the voltage state occurs to the total number of trials. As expected, p_1 approaches 0.5 as P_{10} increases.

In early designs of phase qubits, the junction switched to the voltage state, resulting in energy dissipation. In a later, improved, design[26], the qubits remain in the zero-voltage state (Fig. 5b, c). The qubit junction is embedded in a superconducting loop that is inductively coupled to a SQUID and to a line through which static and pulsed currents can be passed. With appropriately chosen parameters, the potential energy of the qubit displays the two asymmetrical wells shown in Fig. 5c. The states $|0\rangle$ and $|1\rangle$ in the left well are the qubit states; their energy separation and the depth of the well can be controlled by varying the flux in the loop. To read out the state of the phase qubit, a short adiabatic pulse that reduces the depth ΔU of the qubit potential well is applied to the flux bias line. If the qubit is in the state $|1\rangle$, it tunnels rapidly into the right well; in the state $|0\rangle$, no tunnelling occurs. Depending on whether tunnelling occurs, the flux in the qubit loop differs by a single flux quantum, which can easily be detected subsequently by the read-out SQUID. This scheme enables the state of the qubit to be measured rapidly, typically in 5 ns, which is still adiabatic (slow) on the timescale of transitions between the qubit states. Subsequent measurement of the flux in these qubit loops can be made much more slowly.

Time-domain measurements

Spectroscopy is important for establishing that a given qubit is a functional device, and it enables energy-level splitting to be measured as a function of relevant control parameters. But measurements in the

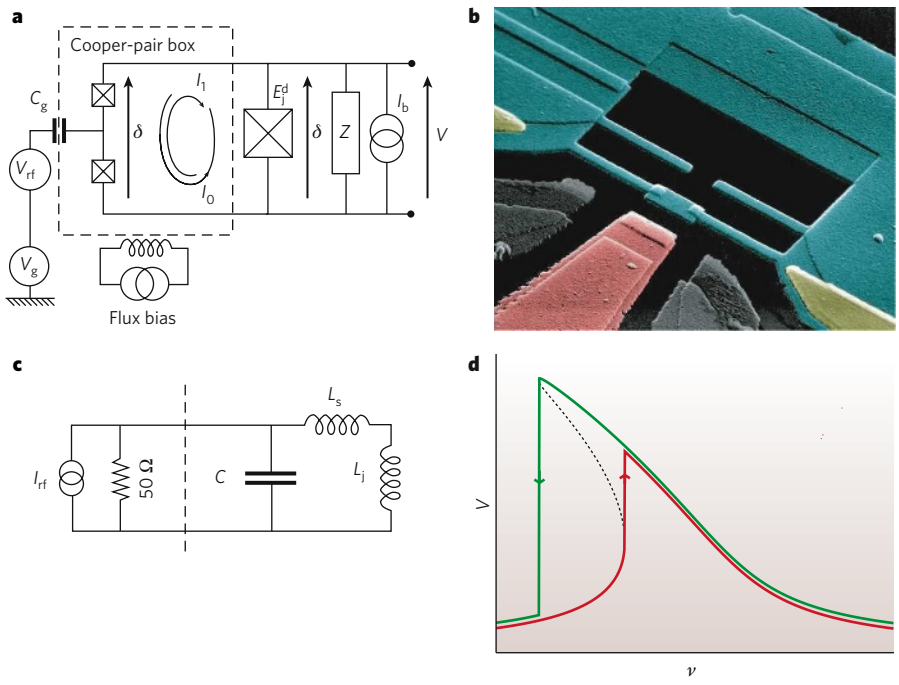


Figure 4 | Quantronium. a, A quantronium circuit is depicted. The Cooper-pair box is connected by way of two Josephson junctions to the detector Josephson junction, which has Josephson energy E^d_j (right), and by way of a capacitor (with gate capacitance C_g) to the static voltage bias V_g and the radio-frequency gate voltage V_{rf} that prepares the state of the Cooper-pair box. The dashed lines enclose the qubit. I_b is the bias current of the detector junction, and Z is an engineered environmental impedance. The flux through the loop formed by the three Josephson junctions is controlled by an external bias circuit. The read-out is the phase δ across the two box junctions, measured by combining the bias current I_b with the circulating loop currents I_0 or I_1. (Panel reproduced, with permission, from ref. 12.) **b**, A micrograph of quantronium is shown. The Cooper-pair box and leads are depicted in blue, and the gate electrode in red. (In gold are normal metal films that are used to remove quasiparticles from the superconducting films.) (Image courtesy of D. Esteve, Commissariat à l'Énergie Atomique, Saclay, France.) **c**, A Josephson bifurcation amplifier (JBA) is depicted. In a JBA, a Josephson junction, represented by the nonlinear inductance L_j, is shunted with a capacitance C via a stray inductance L_S; I_{rf} is the radio-frequency current bias. The dashed line separates the off-chip circuitry (left) from the on-chip circuitry (right). (Panel reproduced, with permission, from ref. 22.) **d**, The response curve (voltage V versus frequency ν) of the JBA driven at high radio-frequency current amplitude at a frequency slightly below resonance is shown, and the hysteresis that results from dynamical bifurcation is indicated (arrows). The red line shows the low-amplitude response of the JBA, and the green line shows the high-amplitude response; the dashed line indicates metastable states.

time domain are also necessary to determine the dynamical behaviour of a qubit. These measurements involve manipulating the state of the qubit by using appropriate microwave pulses — which are also required to implement single-qubit gates for quantum computing. In broad terms, qubits are characterized by two times, named T_1 and T_2 by analogy with nuclear magnetic resonance (NMR) spectroscopy[27]. The relaxation time T_1 is the time required for a qubit to relax from the first excited state to the ground state; this process involves energy loss. The dephasing time T_2 is the time over which the phase difference between two eigenstates becomes randomized. Theoretically, both relaxation and dephasing are described by weak coupling to the quantum noise produced by the environment[27–29]. This approach predicts that energy relaxation arises from fluctuations at the energy-level splitting frequency of the two states in question. The dephasing rate, by contrast, has two contributions:

$$1/T_2 = 1/(2T_1) + 1/\tau_\phi \quad (5)$$

The first contribution arises from the relaxation process, and the second, 'pure dephasing', arises from low-frequency fluctuations with exchange of infinitesimal energy. (The pure dephasing time is τ_ϕ.)

The simplest way to measure relaxation is to irradiate the qubit with microwaves at the frequency corresponding to the energy-level splitting between the ground and first excited states for a time much greater than T_1. After the pulse has been turned off, the qubit has an equal probability of being in either state; the probability p_1 of its being in the excited state $|1\rangle$ subsequently decays with time t as $\exp(-t/T_1)$. Measurements of p_1 as a function of t yield the value of T_1. It should be emphasized that each measurement of p_1 at a given time delay involves a large number of measurements, typically 10^4 or 10^5. T_1 can vary from values of the order of 1 ns to many microseconds.

To understand the various pulse measurements, it is useful to consider the Bloch sphere (Fig. 6a), which enables any arbitrary quantum superposition of the quantum states $|0\rangle$ and $|1\rangle$ to be considered as a vector. The states $|0\rangle$ and $|1\rangle$ point along the positive and negative z axis, respectively. The superpositions $|0\rangle \pm |1\rangle$ lie along the $\pm x$ axes, and the superpositions $|0\rangle \pm i|1\rangle$ along the $\pm y$ axes. Thus, a given point on the surface of the sphere defines a specific superposition of these states.

The Bloch sphere can be used to describe Rabi oscillations in a flux qubit. Microwaves are applied at the energy-level splitting frequency for the qubit for a time τ with the magnetic-field component along the y axis. During the pulse, the state vector rotates in the y–z plane about the x axis with the Rabi frequency ν_R, which is proportional to the microwave amplitude. After time τ, the state vector is at an angle $2\pi\nu_R\tau$ to the z axis. Subsequent measurements of the probability of the qubit being in the state $|0\rangle$ or $|1\rangle$ yield Rabi oscillations as a function of τ. An example is shown in Fig. 6b. Rabi oscillations are a convenient means of calibrating the amplitude of the magnetic-field component of the microwave field that is coupled to the qubit.

In measuring the dephasing time, it is crucial to distinguish T_2 (equation (5)) — an intrinsic timescale for the decoherence of a single qubit — from T_2^*, the result of an ensemble measurement. The ensemble is formed because experiments on a single qubit need to be carried out repeatedly so that sufficiently precise data are acquired. Even though the different measurements are nominally identical, slow fluctuations on the timescale of a single run result in a change in the operating conditions between runs. This reduces the observed coherence time to T_2^* (which is $< T_2$).

T_2^* and T_2 can be measured separately: T_2^*, which includes the effects of low-frequency noise, by using Ramsey fringes[30]; and T_2, by using a spin-echo technique[27], which eliminates certain low-frequency contributions. To observe Ramsey fringes, a $\pi/2$ microwave pulse is first applied at a frequency f_m — with amplitude calibrated from the Rabi oscillations — that tips the qubit state vector into the equatorial (x–y) plane. The vector precesses freely on the Bloch sphere around the static magnetic field B_0, with a magnitude that decreases with time, owing to dephasing. After a variable time delay τ_d, a second $\pi/2$ microwave pulse brings the state vector to a point on the Bloch sphere that depends on both f_m and τ_d. The subsequent measurement of the qubit state projects the vector onto either $|0\rangle$ or $|1\rangle$. Thus, a plot of the switching probability versus τ_d for a given microwave frequency maps out the free evolution of the qubit. For a resonant pulse ($f_m = \nu_{10}$), the free evolution and the microwave pulses are synchronized, and the measurement reveals a coherence amplitude that decays exponentially with characteristic time T_2^*. To map out T_2^* over a larger parameter space, the $\pi/2$ microwave pulses are detuned from ν_{10}. Thus, the pulse and evolution are no longer synchronized, and oscillations — Ramsey fringes — are observed at a frequency $\nu_{Ramsey} = |f_m - \nu_{10}|$ (Fig. 6c).

To remove the slow fluctuations that differentiate T_2^* from T_2, a spin-echo technique, analogous to that used in NMR, can be used. In this technique, a π pulse is applied at the midpoint in time between the two $\pi/2$ pulses. The π pulse flips the qubit state vector to the opposite side of the equatorial plane; therefore, a fluctuation that initially caused the phase to advance now causes it to lag, and vice versa. Thus, at the time of the second $\pi/2$ pulse, the effects of fluctuations that occur on timescales longer than the overall measurement time are (ideally) completely cancelled out. An example is shown in Fig. 6d.

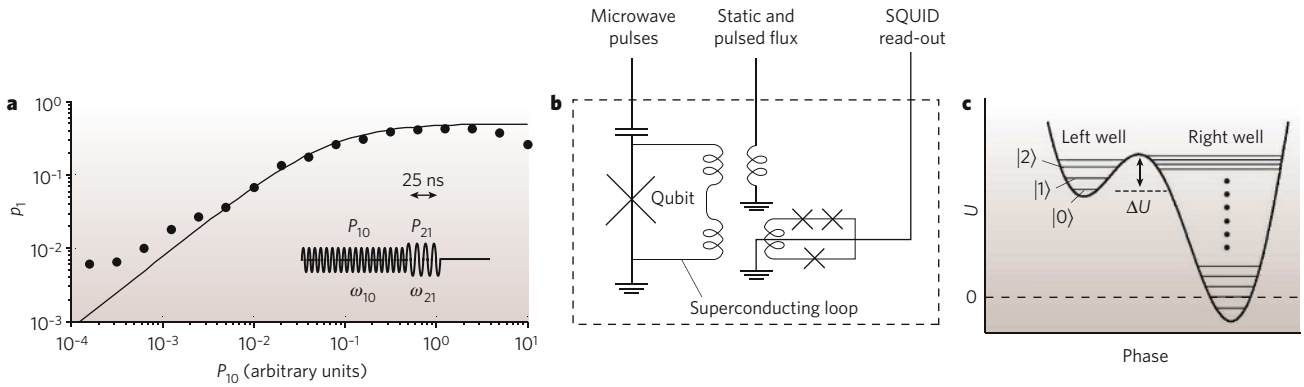


Figure 5 | Phase qubits. a, The filled circles represent the probability p_1 that the phase qubit occupies the first excited state versus microwave power P_{10} at angular frequency ω_{10}. The solid line is the theoretical prediction. The inset shows the pulse sequence; the microwaves at angular frequency ω_{10} equalize the probability that the ground and excited states are occupied, and the microwaves at angular frequency ω_{21} cause the qubit to switch to the voltage state if the first excited state is occupied. (Panel reproduced, with permission, from ref. 13.) **b**, For zero-voltage operation, the Josephson junction of a phase qubit is shunted by a superconducting loop, coupled to a read-out SQUID, that allows static and pulsed fluxes to be applied. The dashed line indicates the components fabricated on the silicon chip, which is maintained at 25 mK. (Panel reproduced, with permission, from ref. 26.) **c**, The asymmetrical double-well potential of a phase qubit is shown. The qubit states are $|0\rangle$ and $|1\rangle$. The state $|2\rangle$ becomes occupied on the application of microwaves at frequency ω_{21} provided that the state $|1\rangle$ is occupied. The state $|3\rangle$, above $|2\rangle$, has no role in the read-out process. Dots in the right well indicate the intervening energy levels. (Panel reproduced, with permission, from ref. 26.)

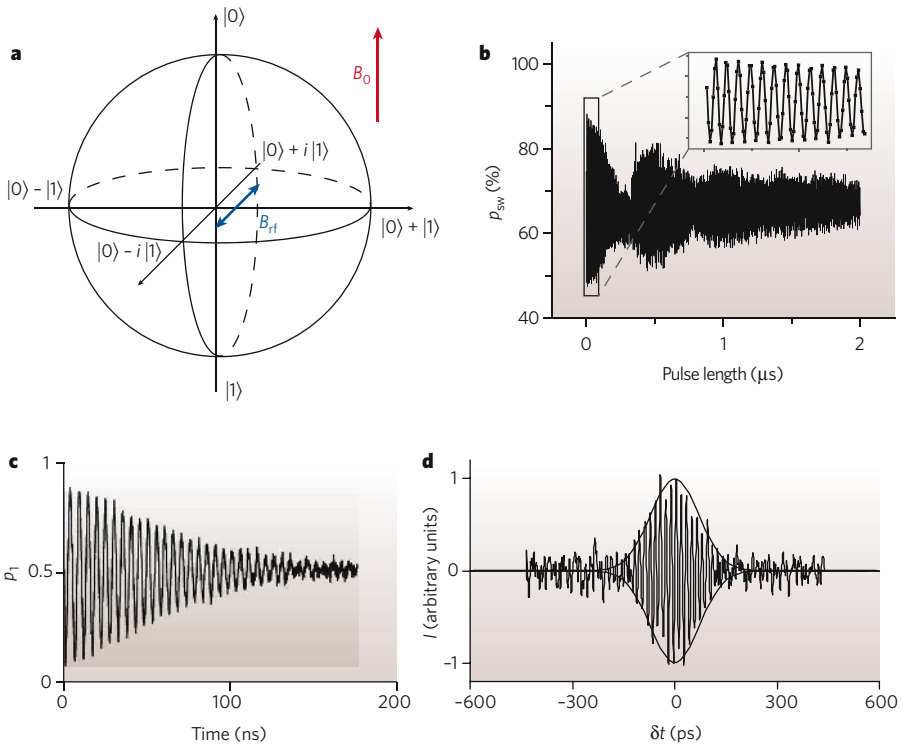


Figure 6 | Qubit manipulation in the time domain. **a**, The Bloch sphere is depicted, with an applied static magnetic field B_0 and a radio-frequency magnetic field B_{rf}. Any given superposition of the six states shown is represented by a unique point on the surface of the sphere. **b**, Rabi oscillations in a flux qubit are shown. The probability p_{sw} that the detector (SQUID) switches to the normal state versus pulse length is shown, and the inset is a magnification of the boxed region, showing that the dense traces are sinusoidal oscillations. As expected, the excited-state population oscillates under resonant driving. (Panel reproduced, with permission, from ref. 40.) **c**, Ramsey fringes in a phase qubit are shown. Coherent oscillations of the switching probability p_1 between two detuned $\pi/2$ pulses is shown as a function of pulse separation. (Panel reproduced, with permission, from ref. 31.) **d**, The charge echo in a Cooper-pair box is shown as a function of the time difference $\delta t = t_1 - t_2$, where t_1 is the time between the initial $\pi/2$ pulse and the π pulse, and t_2 is the time between the π pulse and the second $\pi/2$ pulse. The echo peaks at $\delta t = 0$. (Panel reproduced, with permission, from ref. 39.)

Measuring the times T_1, T_2 and T_2^* provides an important initial characterization of qubit coherence. However, other factors such as pulse inaccuracy, relaxation during measurement and more complex decoherence effects result in measurement errors. A more complete measure of a qubit is fidelity, a single number that represents the difference between the ideal and the actual outcome of the experiment. Determining the fidelity involves quantum-process tomography (a repeated set of state tomographies), which characterizes a quantum-mechanical process for all possible initial states. In a Ramsey-fringe tomography experiment, Matthias Steffen *et al.*[31] found a fidelity of ~80%, where 10% of the loss was attributed to read-out errors and another 10% to pulse-timing uncertainty.

Decoherence

Superconducting qubits are macroscopic, so — along the lines of Schrödinger's cat — they could be expected to be very sensitive to decoherence. In fact, given the unique properties of the superconducting state, careful engineering has led to remarkable increases in decoherence times compared with those of early devices.

Ideally, each type of qubit is described by a single degree of freedom. The central challenge is to eliminate all other degrees of freedom. In broad terms, there are two classes of decohering element: extrinsic and intrinsic. Obvious extrinsic sources include electromagnetic signals from radio and television transmitters; these can generally be eliminated by using careful shielding and enough broadband filters. A more challenging extrinsic source to exclude is the local electromagnetic environment: for example, contributions from the leads that are coupled to read-out devices or are used to apply flux or charge biases. These leads allow great flexibility in control of the system at the expense of considerable coupling to the environment. This issue was recognized in the first proposals of macroscopic quantum coherence and largely motivated the Caldeira–Leggett theory of quantum dissipation[6]. This theory maps any linear dissipation onto a bath of harmonic oscillators. The effects of these oscillators can be calculated from the Johnson–Nyquist noise that is generated by the complex impedance of the environment. In the weak-damping regime, both T_1 and τ_ϕ can be computed directly from the power spectrum of this noise, and then the impedance can be engineered to minimize decoherence[28,29]. The experimental difficulty is to ensure that the complex impedances 'seen' by the qubit are high over a broad bandwidth, for example, 0–10 GHz. It is particularly difficult to avoid resonances over such a broad range of frequencies. Clever engineering has greatly reduced this source of decoherence, but it would be optimistic to consider that this problem has been completely solved.

The main intrinsic limitation on the coherence of superconducting qubits results from low-frequency noise, notably '$1/f$ noise' (in which the spectral density of the noise at low frequency f scales as $1/f^\alpha$, where α is of the order of unity). In the solid state, many $1/f$ noise sources are well described by the Dutta–Horn model as arising from a uniform distribution of two-state defects[32]. Each defect produces random telegraph noise, and a superposition of such uncorrelated processes leads to a $1/f$ power spectrum. There are three recognized sources of $1/f$ noise. The first is critical-current fluctuations, which arise from fluctuations in the transparency of the junction caused by the trapping and untrapping of electrons in the tunnel barrier[33]. All superconducting qubits are subject to dephasing by this mechanism. The slow fluctuations modulate energy-level splitting, even at the degeneracy point, so each measurement is made on a qubit with a slightly different frequency. The resultant phase errors lead to decoherence.

The second source of $1/f$ noise is charge fluctuations, which arise from the hopping of electrons between traps on the surface of the superconducting film or the surface of the substrate. This motion induces charges onto the surface of nearby superconductors. This decoherence mechanism is particularly problematic for charge qubits, except at the degeneracy point, where the qubits are (to first order) insensitive. If the value of E_c/E_j increases, however, the energy bands (Fig. 3c) become flatter, and the qubit is correspondingly less sensitive to charge noise away from the degeneracy point. This mechanism underlies the substantially increased values of T_2 in the transmon[20].

The third source of $1/f$ noise is magnetic-flux fluctuations. Although such fluctuations were first characterized more than 20 years ago[34], the mechanism by which these occur remained obscure until recently. It is now thought that flux noise arises from the fluctuations of unpaired electron spins on the surface of the superconductor or substrate[35,36], but the details of the mechanism remain controversial. Flux noise causes decoherence in flux qubits, except at the degeneracy point, as well as in phase qubits, which have no degeneracy point. The increased value of T_2 in quantronium results from its insensitivity to both flux noise and charge noise at the double degeneracy point.

Table 1 | Highest reported values of T_1, T_2^* and T_2

Qubit	T_1 (μs)	T_2^* (μs)	T_2 (μs)	Source
Flux	4.6	1.2	9.6	Y. Nakamura, personal communication
Charge	2.0	2.0	2.0	ref. 77
Phase	0.5	0.3	0.5	J. Martinis, personal communication

In general, all three low-frequency processes lead to decoherence. They do not contribute to relaxation because this process requires an exchange of energy with the environment at the energy-level splitting frequency of the qubit, which is typically in the gigahertz range. However, there is strong evidence that charge fluctuations are associated with the high-frequency resonators that have been observed, in particular, in phase qubits[37]. Improvements in the quality of the oxide layers that are used in the junctions and capacitors have resulted in large reductions in the concentration of these high-frequency resonators[38].

The strategy of operating a qubit at the optimum point, which was first carried out with quantronium but is now applied to all types of superconducting qubit (except for phase qubits), has been successful at increasing phase-coherence times by large factors. Further substantial improvements have resulted from the use of charge- or flux-echo techniques[39,40]. In NMR, the spin-echo technique removes the inhomogeneous broadening that is associated with, for example, variations in magnetic field, and hence in the NMR frequency, over the sample. In the case of qubits, the variation is in the qubit energy-level splitting frequency from measurement to measurement. For some qubits, using a combination of echo techniques and optimum point operation has eliminated pure dephasing, so decoherence is limited by energy relaxation ($T_2^* = 2T_1$). In general, however, the mechanisms that limit T_1 are unknown, although resonators that are associated with defects may be responsible[36,41]. The highest reported values of T_1, T_2^* and T_2 are listed in Table 1.

Coupled qubits

An exceedingly attractive and unique feature of solid-state qubits in general and superconducting qubits in particular is that schemes can be implemented that both couple them strongly to each other and turn off their interaction *in situ* by purely electronic means. Because the coupling of qubits is central to the architecture of quantum computers, this subject has attracted much attention, in terms of both theory and experiment. In this section, we illustrate the principles of coupled qubits in terms of flux qubits and refer to analogous schemes for other superconducting qubits.

Because the flux qubit is a magnetic dipole, two neighbouring flux qubits are coupled by magnetic dipole–dipole interactions. The coupling strength can be increased by having the two qubits use a common line. Even stronger coupling can be achieved by including a Josephson junction in this line to increase the line's self-inductance (equation (6), Box 1). In the case of charge and phase qubits, nearest-neighbour interactions are mediated by capacitors rather than inductors. Fixed interaction has been implemented for flux, charge and phase qubits[42–45]. These experiments show the energy levels that are expected for the superposition of two pseudospin states: namely, a ground state and three excited states; the first and second excited states may be degenerate. The entanglement of these states for two phase qubits has been shown explicitly by means of quantum-state tomography[46]. The most general description (including all imperfections) of the qubit state based on the four basis states of the coupled qubits is a four-by-four array known as a density matrix. Steffen *et al.*[46] carried out a measurement of the density matrix; they prepared a system in a particular entangled state and showed that only the correct four matrix elements were non-zero — and that their magnitude was in good agreement with theory. This experiment is a proof-of-principle demonstration of a basic function required for a quantum computer. Simple quantum gates have also been demonstrated[47,48].

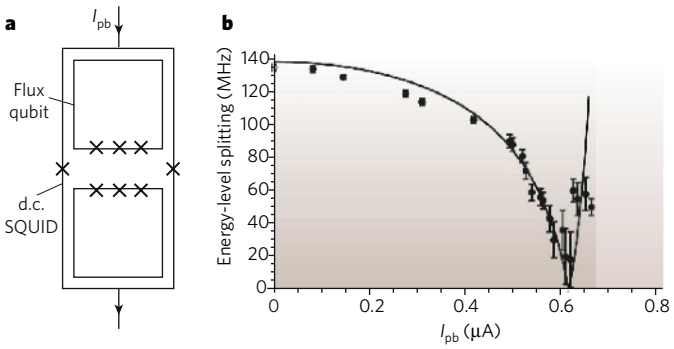


Figure 7 | Controllably coupled flux qubits. a, Two flux qubits are shown surrounded by a d.c. SQUID. The qubit coupling strength is controlled by the pulsed bias current I_{pb} that is applied to the d.c. SQUID before measuring the energy-level splitting between the states $|1\rangle$ and $|2\rangle$. **b**, The filled circles show the measured energy-level splitting of the two coupled flux qubits plotted against I_{pb}. The solid line is the theoretical prediction, fitted for I_{pb}; there are no fitted parameters for the energy-level splitting. Error bars, ±1σ. (Panels reproduced, with permission, from ref. 50.)

Two flux qubits can be coupled by flux transformers — in essence a closed loop of superconductor surrounding the qubits — enabling their interaction to be mediated over longer distances. Because the superconducting loop conserves magnetic flux, a change in the state of one qubit induces a circulating current in the loop and hence a flux in the other qubit. Flux transformers that contain Josephson junctions enable the interaction of qubits to be turned on and off *in situ*. One such device consists of a d.c. SQUID surrounding two flux qubits[49] (Fig. 7a). The inductance between the two qubits has two components: that of the direct coupling between the qubits, and that of the coupling through the SQUID. For certain values of applied bias current (below the critical current) and flux, the self-inductance of the SQUID becomes negative, so the sign of its coupling to the two qubits opposes that of the direct coupling. By choosing parameters appropriately, the inductance of the coupled qubits can be designed to be zero or even have its sign reversed. This scheme has been implemented by establishing the values of SQUID flux and bias current and then using microwave manipulation and measuring the energy-level splitting of the first and second excited states[50] (Fig. 7b). A related design — tunable flux–flux coupling mediated by an off-resonant qubit — has been demonstrated[51], and tunable capacitors have been proposed for charge qubits[52].

Another approach to variable coupling is to fix the coupling strength geometrically and tune it by frequency selection. As an example, we consider two magnetically coupled flux qubits biased at their degeneracy points. If each qubit is in a superposition of eigenstates, then its magnetic flux oscillates and the coupling averages to zero — unless both qubits oscillate at the same frequency, in which case the qubits are coupled. This phenomenon is analogous to the case of two pendulums coupled by a weak spring. Even if the coupling is extremely weak, the pendulums will be coupled if they oscillate in antiphase at exactly the same frequency.

Implementing this scheme is particularly straightforward for two phase qubits because their frequencies can readily be brought in and out of resonance by adjusting the bias currents[37]. For other types of qubit, the frequency at the degeneracy point is set by the as-fabricated parameters, so it is inevitable that there will be variability between qubits. As a result, if the frequency difference is larger than the coupling strength, the qubit–qubit interaction cancels out at the degeneracy point. Several pulse sequences have been proposed to overcome this limitation[53–55], none of which has been convincingly demonstrated as yet. The two-qubit gate demonstrations were all carried out away from the optimum point, where the frequencies can readily be matched.

On the basis of these coupling schemes, several architectures have been proposed for scaling up from two qubits to a quantum computer. The central idea of most proposals is to couple all qubits to a long central coupling element, a 'quantum bus'[56,57] (Fig. 8), and to use frequency selection to determine which qubits can be coupled[56–60]. This scheme has been experimentally demonstrated. As couplers become longer, they become transmission lines that have electromagnetic modes. For example, two

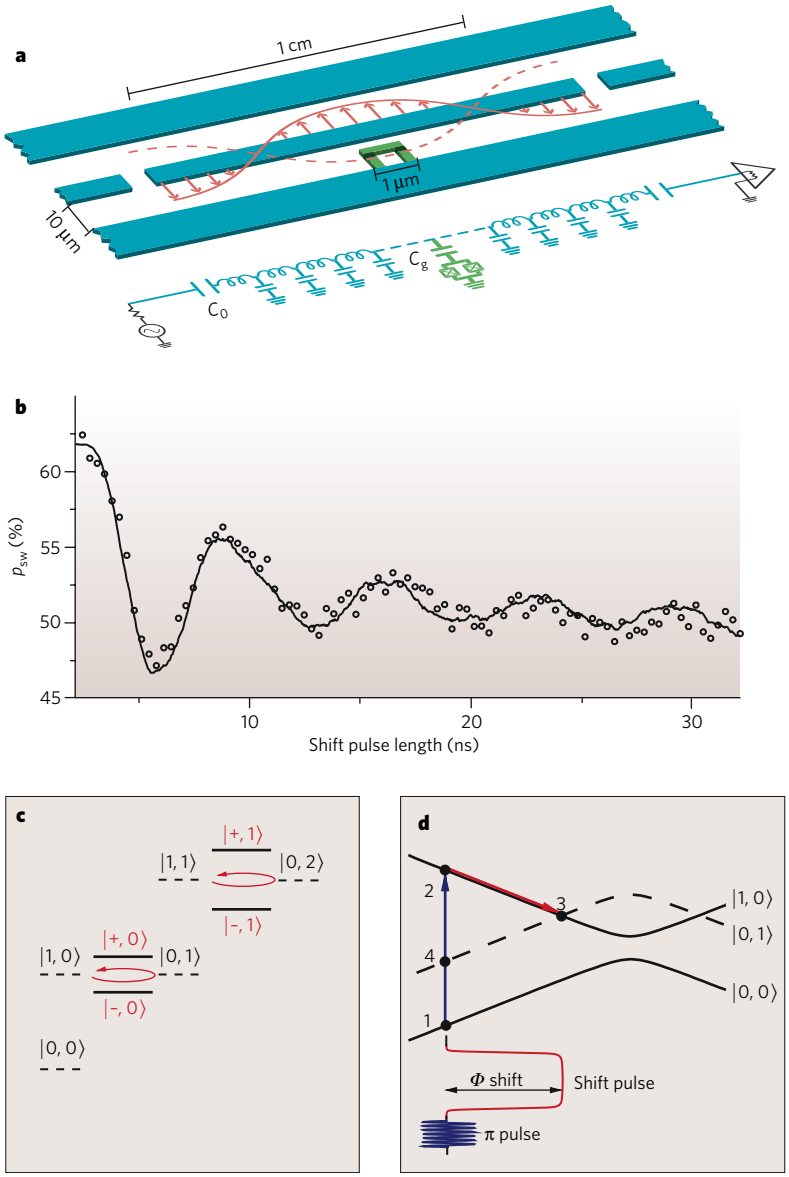


Figure 8 | Circuit QED. a, The upper part of the panel depicts a microstrip cavity (blue) that contains a charge qubit (green) placed at an antinode of the electric field. The microstripline can be used as a quantum bus. The lower part depicts this circuit in a lumped circuit representation. (Panel reproduced, with permission, from ref. 59.) C_0 is the capacitance of the coupling capacitor to the measurement electronics, and C_g is the capacitance of the coupling capacitor to the charge qubit. **b**, The open circles show the measured vacuum Rabi oscillations of a flux qubit coupled to a lumped resonator. The solid curve is a fit to the data. (Panel reproduced, with permission, from ref. 68.) **c**, An energy ladder of qubit ground and excited states combined with photon number n, $|0, n\rangle$ and $|1, n\rangle$ (dashed lines), is shown. With the cavity in resonance with the qubits, the states with zero photons split into linear combinations $|\pm, 0\rangle$ (solid lines), with an energy-level splitting g, and the states with one photon split into linear combinations $|\pm, 1\rangle$, with an energy-level splitting $\sqrt{2}g$. The red arrows indicate that if the system is initially in one of the states represented by dashed lines, it will perform Rabi oscillations between the qubit and the cavity. (Panel modified, with permission, from ref. 68.) **d**, An energy-band diagram (solid and dashed black lines) is shown as a function of applied flux for the measurement scheme that led to the results in **b**. The measurement pulse (π pulse) forces the system from the ground state (point 1) into a state with an excited qubit (point 2) (depicted in blue), which then puts the qubit and the cavity into resonance at point 3 (depicted in red). After the vacuum Rabi oscillation occurs, the system returns to point 2 or makes a coherent transition to point 4, where the qubit excitation is converted to a cavity photon. (Panel modified, with permission, from ref. 68.)

qubits have been coupled by placing them at the antinodes of a standing wave on a stripline[59–62]. Coupling between specific pairs of qubits can result in a scalable architecture[63]. By first coupling a qubit to the standing-wave mode using frequency selection, a photon is excited and then stored after decoupling. Subsequently, a second qubit is coupled to the mode, and the photon transfers the quantum state to the second qubit.

Architectures for adiabatic quantum computers are the subject of intense research. Adiabatic quantum computing encodes the solution to a hard problem in the ground state of a qubit system and uses quantum physics to prepare that ground state efficiently. The ground state of a four-qubit system with tunable interactions has been mapped out[64]. It should, however, be noted that there is no proof that an adiabatic quantum computer will be faster than a classical computer.

Quantum optics on a chip

An important new direction in superconducting qubit research is based on analogy between superconducting circuits and the fields of atomic physics and quantum optics. So far, we have described only qubits as quantum objects, and the control fields and read-out signals have been treated as classical variables. Circuit QED, by contrast, addresses the quantum behaviour of the electromagnetic field, such as that of single photons. In previous sections, the discussion refers to a quantum field in a coherent state in the limit of large numbers of photons.

The key requirement for reaching the quantum limit of the electromagnetic field is that the zero-point fluctuation of a single mode — measured by the root mean square of the electric field, $E_{rms} = \sqrt{\langle E^2 \rangle_{vacuum}}$ — be strong enough to have an appreciable coupling strength $g = dE_{rms}$ to the qubit electric dipole moment d. This requirement is met by increasing the amplitude of the field by creating a standing wave in a resonator and placing the qubit at one of the antinodes[59] (Fig. 8a). The resonator can be either a microstripline — an on-chip wave guide for microwaves — or a lumped circuit. In the first experiment[65], the resonator was tunable. The physics is closely related to cavity QED[66], in which atoms couple to an optical field confined between two mirrors. A key difference is that in circuit QED, the 'atom' (that is, the superconducting qubit) does not move inside the cavity, so the 'atom'–field interaction has time to act without losing the 'atom'. Together with the fact that $g/\hbar$ is larger than the rate of photon loss from the cavity, this difference allows the strong coupling limit of QED to be achieved in a relatively straightforward manner. The underlying reasons are that g is proportional to d (which, for a Cooper-pair box, is large, about 10^4 atomic units) and that E_{rms} is also large because of the increase in the electromagnetic field in the one-dimensional stripline.

Circuit QED can be operated in two distinct strong-coupling limits: the resonant regime, and the off-resonant dispersive regime. In the resonant regime, the qubit energy-level splitting is in resonance with the cavity

frequency. In this regime, the combined states of the qubit and cavity can be written in the form |qubit state, photon number⟩. On resonance, the qubit and cavity can exchange excitations without losing energy: that is, the energy of $|1, n\rangle$ is equal to the energy of $|0, n+1\rangle$. The eigenstates of the system are thus superpositions of the form $|\pm, n\rangle = |1, n\rangle \pm |0, n+1\rangle$, with energies split by $g\sqrt{n}$, leading to the energy spectrum shown in Fig. 8c. This has a striking consequence: suppose that initially the qubit energy is not in resonance with the cavity (so the two are decoupled) and that the qubit is put into an excited state while the cavity is left in its vacuum state. When the qubit and the cavity in that state are suddenly coupled by using the procedure shown in Fig. 8d, the original state ceases to be an eigenstate and, instead, becomes an equal superposition of $|+, 0\rangle$ and $|-, 0\rangle$. After a time t, these acquire a relative phase of $gt/\hbar$ and manifest themselves as a coherent oscillation between $|1, 0\rangle$ and $|0, 1\rangle$, even though initially there was no photon in the cavity. These vacuum Rabi oscillations have been shown spectroscopically[67] and in the time domain[68] (Fig. 8b).

The second case is the off-resonant dispersive regime. In this case, the qubit and cavity eigenstates are not entangled, and the two systems cannot share excitations. The mutual energies, however, are still correlated, because the energy-level splitting of the qubit depends on the cavity state, and vice versa. Consequently, the cavity can be used to read out the qubit and to couple qubits to each other[59].

Circuit QED has been highly successful. So far, experimental progress has included attaining the strong coupling limit[67], mapping out the discrete nature of the quantized field[69], generating single photons[70] and coupling qubits using a bus[61,62]. These developments are leading to flexible quantum optics on a chip and open the door to a new domain of mesoscopic physics. Scalable architectures for quantum computers based on circuit QED have been proposed[62].

These ideas have led to the recent demonstration of a superconducting qubit laser. The 'atom' — a charge qubit — is weakly coupled to a second lead. In appropriate bias conditions, a cyclic process takes place: Cooper pairs that enter the box are broken into two quasiparticles, which exit through the second lead. This cycle results in a significant overpopulation of the first excited qubit state compared with the ground state — that is, a population inversion — and the generation of a laser action[71].

Studies in atomic physics have produced superb techniques for actively cooling atoms. Because superconducting qubits operate at millikelvin temperatures, it might be thought that further cooling is unnecessary. But both the preparation of a high-fidelity initial state and the supply of qubits initialized to the ground state for error correction can be facilitated by active cooling. Cooling to 3 mK from an initial temperature of 400 mK has been achieved by exciting the population of the excited state of a flux qubit to a higher excited state that is delocalized in a double-well potential, and then allowing the qubit to relax to the ground state[72]. 'Sisyphus cooling' has also been demonstrated[73]: in this cooling protocol, the energy that is supplied to the qubit from the heat bath is cyclically removed by the magnetic component of a suitably tailored microwave field.

Outlook

Quantum computing is a huge driving force for technological innovation. Since macroscopic quantum coherence was shown, the progress in the design and operation of superconducting qubits has been remarkable. There is now a rich variety of devices that contain the three qubit types, either separately or in combination. Decoherence times have been increased from ~1 ns to ~10 μs, and single-shot and QND read-outs are close to being achieved. So, what challenges and prospects now lie ahead?

On a fundamental level, the next benchmark is to verify a violation of Bell's inequality[74]. This inequality, which involves the outcomes of a combination of two-qubit measurements, is obeyed for any local theory but is violated for truly non-local physics such as quantum mechanics. A variation is the Leggett–Garg inequality[75], which relates to temporal correlations rather than to two-qubit correlations. One important aspect of quantum mechanics — entanglement — has been shown for superconducting qubits[46], but the testing of whether Bell's inequality is violated poses formidable technological challenges, particularly with respect to the fidelity of the measurement and the elimination of cross-talk. To make a Bell test convincing, the interaction between qubits needs to be switched off very accurately so that measurements are truly independent. An even more convincing test would involve a true space-like separation: that is, measuring the read-out of two qubits in such a short time that no signal has been able to travel between them at the speed of light. Given the confines of a dilution refrigerator, however, it seems that it will not be possible to test superconducting qubits in this way.

Another important experiment involving entanglement will be to investigate whether teleportation of a state occurs[76]: that is, the transfer of a quantum state inside an entangled pair of states.

On the path to quantum computing, superconducting qubits are clearly among the most promising candidates. Nevertheless, the path is long, and there are quantitative technological obstacles to be overcome, notably increasing the decoherence time and improving the fidelity of the read-out. The key benchmark will be to demonstrate simple error correction. To achieve these grand goals will require technological progress, not the least in the elimination — or at least the reduction — of low-frequency noise. Two-qubit coherence — in particular, the question of whether noise processes are correlated between qubits — is largely unexplored.

Will there ever be a superconducting quantum computer? This question cannot be answered today. The error thresholds discussed in fault-tolerance research — 1 error in 10,000 operations being a typical, but by no means universal, benchmark — are daunting. However, fault-tolerance research is an evolving field, and the computational protocols that have been discussed so far (which minimize the number of physical qubits and interactions required for a given algorithm) might not be best suited for superconducting qubits. Promising alternatives might be error-correction models with more generous thresholds, topological computing or other alternative computational models. Adiabatic quantum computing could also be an alternative if it is proved to be faster than classical computing. While addressing these issues, researchers are also likely to gain further insight into many physical properties and processes. ■

1. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of superconductivity. *Phys. Rev.* **108,** 1175–1204 (1957).
2. Tinkham, M. *Introduction to Superconductivity* (McGraw-Hill, New York, 1996).
3. Devoret, M. H. in *Quantum Fluctuations: Les Houches Session LXIII* (eds Reynaud, S., Giacobino, E. & David, F.) 351–386 (Elsevier, Amsterdam, 1997).
4. Voss, R. F. & Webb, R. A. Macroscopic quantum tunneling in 1-μm Nb Josephson junctions. *Phys. Rev. Lett.* **47,** 265–268 (1981).
5. Devoret, M. H., Martinis, J. M. & Clarke, J. Measurements of macroscopic quantum tunneling out of the zero-voltage state of a current-biased Josephson junction. *Phys. Rev. Lett.* **55,** 1908–1911 (1985).
6. Caldeira, A. O. & Leggett, A. J. Quantum tunneling in a dissipative system. *Ann. Phys. (NY)* **149,** 374–456 (1983).
7. Martinis, J. M., Devoret, M. H. & Clarke, J. Energy-level quantization in the zero-voltage state of a current-biased Josephson junction. *Phys. Rev. Lett.* **55,** 1543–1546 (1985).
8. Leggett, A. J. in *Chance and Matter: Les Houches Session XLVI* (eds Souletie, J., Vannimenus, J. & Stora, R.) 395–506 (Elsevier, Amsterdam, 1987).
9. Nakamura, Y., Chen, C. D. & Tsai, J. S. Spectroscopy of energy-level splitting between two macroscopic quantum states of charge coherently superposed by Josephson coupling. *Phys. Rev. Lett.* **79,** 2328–2331 (1997).
10. Friedman, J. R., Patel, V., Chen, W.,Tolpygo, S. K. & Lukens, J. E. Quantum superpositions of distinct macroscopic states. *Nature* **406,** 43–46 (2000).
11. van der Wal, C. H. *et al.* Quantum superpositions of macroscopic persistent current. *Science* **290,** 773–776 (2000).
12. Vion, D. *et al.* Manipulating the quantum state of an electrical circuit. *Science* **296,** 886–889 (2002).
13. Martinis, J. M., Nam, S., Aumentado, J. & Urbina, C. Rabi oscillations in a large Josephson-junction qubit. *Phys. Rev. Lett.* **89,** 117901 (2002).
14. *The SQUID Handbook: Fundamentals and Technology of SQUIDs and SQUID Systems* Vol. I (eds Clarke, J. & Braginski, A. I.) (Wiley, Weinheim, 2004).
15. Wilhelm, F. K. *et al.* Macroscopic quantum superposition of current states in a Josephson junction loop. *Usp. Fiz. Nauk* **44** (suppl. 171), 117–121 (2001).
16. Devoret, M. H. & Schoelkopf, R., Amplifying quantum signals with the single-electron transistor. *Nature* **406,** 1039–1046 (2000).
17. Schoelkopf, R. J., Wahlgren, P., Kozhevnikov, A. A., Delsing, P. & Prober, D. E. The radio-frequency single-electron transistor (RF-SET): a fast and ultrasensitive electrometer. *Science* **280,** 1238–1242 (1998).
18. Bouchiat, V., Vion, D., Joyez, P., Esteve, D. & Devoret, M. H. Quantum coherence with a single Cooper pair. *Physica Scripta* **T76,** 165–170 (1998).
19. Nakamura, Y., Pashkin, Y. A. & Tsai, J. S., Coherent control of macroscopic quantum states in a single-Cooper-pair box. *Nature* **398,** 786–788 (1999).

20. Koch, J. *et al.* Charge-insensitive qubit design derived from the Cooper pair box. *Phys. Rev. A* **76,** 042319 (2007).
21. Ithier, G. *et al.* Decoherence in a superconducting quantum bit circuit *Phys. Rev. B* **72,** 134519 (2005).
22. Siddiqi, I. *et al.* Direct observation of dynamical bifurcation between two driven oscillation states of a Josephson junction. *Phys. Rev. Lett.* **94,** 027005 (2005).
23. Braginsky, V. B., Khalili, F. Y. & Thorne, K. S. *Quantum Measurement* (Cambridge Univ. Press, Cambridge, UK, 1995).
24. Lupaşcu, A. *et al.* Quantum non-demolition measurement of a superconducting two-level system. *Nature Phys.* **3,** 119–125 (2007).
25. Grajcar, M. *et al.* Low-frequency measurement of the tunneling amplitude in a flux qubit. *Phys. Rev. B* **69,** 060501 (2004).
26. Cooper, K. B. *et al.* Observation of quantum oscillations between a Josephson phase qubit and a microscopic resonator using fast readout. *Phys. Rev. Lett.* **93,** 180401 (2004).
27. Slichter C. P. *Principles of Nuclear Magnetic Resonance* 3rd edn (Springer, New York, 1990).
28. Makhlin, Y., Schön, G. & Shnirman, A. Quantum-state engineering with Josephson-junction devices. *Rev. Mod. Phys.* **73,** 357–400 (2001).
29. Wilhelm, F. K., Hartmann, U., Storcz, M. J. & Geller, M. R. in *Manipulating Quantum Coherence in Solid State Systems* (eds Flatté, M. E. & Tifrea, I.) 195–233 (Springer, Dordrecht, 2007).
30. Ramsey, N. F. A molecular beam resonance method with separated oscillating fields. *Phys. Rev.* **78,** 695–699 (1950).
31. Steffen, M. *et al.* State tomography of capacitively shunted phase qubits with high fidelity. *Phys. Rev. Lett.* **97,** 050502 (2006).
32. Dutta P. & Horn P. M. Low frequency fluctuations in solids: $1/f$ noise. *Rev. Mod. Phys.* **53,** 497–516 (1981).
33. van Harlingen, D. J. *et al.* Decoherence in Josephson-junction qubits due to critical-current fluctuations. *Phys. Rev. B* **70,** 064517 (2004).
34. Wellstood, F. C., Urbina, C. & Clarke, J. Low-frequency noise in dc superconducting quantum interference devices below 1 K. *Appl. Phys. Lett.* **50,** 772–774 (1987).
35. Koch, R. H., DiVincenzo, D. P. & Clarke, J. Model for $1/f$ flux noise in SQUIDs and qubits. *Phys. Rev. Lett.* **98,** 267003 (2007).
36. Faoro, L. & Ioffe, L. B. Microscopic origin of low-frequency flux noise in Josephson circuits. Preprint at <http://arxiv.org/abs/0712.2834> (2007).
37. Simmonds, R. W., Lang, K. M., Hite, D. A., Pappas, D. P. & Martinis, J. M. Decoherence in Josephson qubits from junction resonances. *Phys. Rev. Lett.* **93,** 077033 (2004).
38. Martinis, J. M. *et al.* Decoherence in Josephson qubits from dielectric loss. *Phys. Rev. Lett.* **95,** 210503 (2005).
39. Nakamura, Y., Pashkin, Y. A., Yamamoto, T. & Tsai, J. S. Charge echo in a Cooper-pair box. *Phys. Rev. Lett.* **88,** 047901 (2002).
40. Bertet, P. *et al.* Relaxation and dephasing in a flux-qubit. *Phys. Rev. Lett.* **95,** 257002 (2005).
41. Astafiev, O., Pashkin, Y. A., Nakamura, Y., Yamamoto, T. & Tsai, J. S. Quantum noise in the Josephson charge qubit. *Phys. Rev. Lett.* **93,** 267007 (2004).
42. Berkley, A.J. *et al.* Entangled macroscopic quantum states in two superconducting qubits. *Science* **300,** 1548–1550 (2003).
43. Majer, J. B., Paauw, F. G., ter Haar, A. C. J., Harmans, C. J. P. M. & Mooij, J. E. Spectroscopy of coupled flux qubits. *Phys. Rev. Lett.* **94,** 090501 (2005).
44. Pashkin, Y. A. *et al.* Quantum oscillations in two coupled charge qubits. *Nature* **421,** 823–826 (2003).
45. McDermott, R. *et al.* Simultaneous state measurement of coupled Josephson phase qubits. *Science* **307,** 1299–1302 (2005).
46. Steffen, M. *et al.* Measurement of the entanglement of two superconducting qubits via state tomography. *Science* **313,** 1423–1425 (2006).
47. Yamamoto, Y., Pashkin, Y. A., Astafiev, O., Nakamura, Y. & Tsai, J. S. Demonstration of conditional gate operation using superconducting charge qubits. *Nature* **425,** 941–944 (2003).
48. Plantenberg, J. H., de Groot, P. C., Harmans, C. J. & Mooij, J. E. Demonstration of controlled-NOT quantum gates on a pair of superconducting quantum bits. *Nature* **447,** 836–839 (2007).
49. Plourde, B. L. T. *et al.* Entangling flux qubits with a bipolar dynamic inductance. *Phys. Rev. B* **70,** 140501 (2004).
50. Hime, T. *et al.* Solid-state qubits with current-controlled coupling. *Science* **314,** 1427–1429 (2006).
51. Niskanen, A. O. *et al.* Quantum coherent tunable coupling of superconducting qubits. *Science* **316,** 723–726 (2007).
52. Averin, D. V. & Bruder, C. Variable electrostatic transformer controllable coupling of two charge qubits. *Phys. Rev. Lett.* **91,** 057003 (2003).
53. Bertet, P., Harmans, C. J. P.M. & Mooij, J. E. Parametic coupling for superconducting qubits. *Phys. Rev. B* **73,** 064512 (2006).
54. Rigetti, C., Blais, A. & Devoret, M. Protocol for universal gates in optimally biased superconducting qubits. *Phys. Rev. Lett.* **94,** 240502 (2005).
55. Liu, Y.-x., Wei, L. F., Tsai, J. S. & Nori, F. Controllable coupling between flux qubits. *Phys. Rev. Lett.* **96,** 067003 (2006).
56. Makhlin, Y., Schön, G. & Shnirman, A. Josephson-junction qubits with controlled couplings. *Nature* **398,** 305–307 (1999).
57. Wei, L. F., Liu, Y.-x. & Nori, F. Quantum computation with Josephson qubits using a current-biased information bus. *Phys. Rev. B* **71,** 134506 (2005).
58. Lantz, J., Wallquist, M., Shumeiko, V. S. & Wendin, G. Josephson junction qubit network with current-controlled interaction. *Phys. Rev. B* **70,** 140507 (2004).
59. Blais, A., Huang, R.-S., Wallraff, A., Girvin, S. M. & Schoelkopf, R. J. Cavity quantum electrodynamics for superconducting electrical circuits: an architecture for quantum computation. *Phys. Rev. A* **69,** 062320 (2004).
60. Helmer, F. *et al.* Two-dimensional cavity grid for scalable quantum computation with superconducting circuits. Preprint at <http://arxiv.org/abs/0706.3625> (2007).
61. Majer, J. B. *et al.* Coupling superconducting qubits via a cavity bus. *Nature* **449,** 443–447 (2007).
62. Sillanpaa, M. A., Park, J. I. & Simmonds, R. W. Coherent quantum state storage and transfer between two phase qubits via a resonant cavity. *Nature* **449,** 438–442 (2007).
63. Fowler, A. G. *et al.* Long-range coupling and scalable architecture for superconducting flux qubits. *Phys. Rev. B* **76,** 174507 (2007).
64. Grajcar, M. *et al.* Four-qubit device with mixed couplings. *Phys. Rev. Lett.* **96,** 047006 (2006).
65. Devoret, M. H. *et al.* in *Quantum Tunneling in Condensed Media* (eds Kagan, Y. & Leggett, A. J.) 313–345 (Elsevier, Amsterdam, 1992).
66. Haroche, S. & Kleppner, D. Cavity quantum electrodynamics. *Phys. Today* **42,** 24–26 (1989).
67. Wallraff, A. *et al.* Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431,** 162–167 (2004).
68. Johansson, J. *et al.* Vacuum Rabi oscillations in a macroscopic superconducting qubit *LC* oscillator system. *Phys. Rev. Lett.* **96,** 127006 (2006).
69. Schuster, D. I. *et al.* Resolving photon number states in a superconducting circuit. *Nature* **445,** 515–518 (2007).
70. Houck, A. A. *et al.* Generating single microwave photons in a circuit. *Nature* **449,** 443–447 (2007).
71. Astafiev, O. *et al.* Single artificial-atom lasing. *Nature* **449,** 588–590 (2007).
72. Valenzuela, S. O. *et al.* Microwave-induced cooling of a superconducting qubit. *Science* **314,** 1589–1592 (2006).
73. Gracjar, M. *et al.* Sisyphus damping and amplification by a superconducting qubit. Preprint at <http://arxiv.org/abs/0708.0665> (2007).
74. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23,** 880–884 (1969).
75. Leggett, A. J. & Garg, A. Quantum mechanics versus macroscopic realism: is the flux there when nobody looks? *Phys. Rev. Lett.* **54,** 857–860 (1985).
76. Bennett, C. H. *et al.* Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. *Phys. Rev. Lett.* **70** 1895–1899 (1993).
77. Schreier, J. A. *et al.* Suppressing charge noise decoherence in superconducting charge qubits. *Phys. Rev. B* **77,** 180502 (2008).
78. Duty, T., Gunnarson, D., Bladh, K. & Delsing, P. Coherent dynamics of a Josephson charge qubit. *Phys. Rev. B* **69**, 140503 (2004).
79. Dolan, G. J. Offset masks for liftoff photoprocessing. *Appl. Phys. Lett.* **31,** 337–339 (1977).

Acknowledgements Our work is supported by the US Department of Energy (Division of Materials Sciences and Engineering, in the Office of Basic Energy Sciences) (J.C.), and by the Natural Sciences and Engineering Research Council of Canada, QuantumWorks and EuroSQIP (F.K.W.).

Author Information Reprints and permissions information is available at npg.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to the authors (jclarke@berkeley.edu; fwilhelm@iqc.ca).

Coherent manipulation of single spins in semiconductors

Ronald Hanson[1] & David D. Awschalom[2]

During the past few years, researchers have gained unprecedented control over spins in the solid state. What was considered almost impossible a decade ago, in both conceptual and practical terms, is now a reality: single spins can be isolated, initialized, coherently manipulated and read out using both electrical and optical techniques. Progress has been made towards full control of the quantum states of single and coupled spins in a variety of semiconductors and nanostructures, and towards understanding the mechanisms through which spins lose coherence in these systems. These abilities will allow pioneering investigations of fundamental quantum-mechanical processes and provide pathways towards applications in quantum information processing.

In the past few decades, the application of nuclear magnetic resonance and electron spin resonance to large spin ensembles has yielded substantial information on spin dynamics in semiconductors. Experimental advances since the 1990s have allowed researchers to increase their control over single charges, providing a pathway for studies of single spins. Early experiments on single spins confined in semiconductor quantum dots highlighted the opportunity for controlling individual quantum states in a solid.

When quantum information processing became a realistic prospect in the late 1990s, Daniel Loss and David DiVincenzo proposed a quantum computing scheme based on spins in quantum dots[1], and Bruce Kane developed a proposal for a silicon-based quantum computer[2]. It was apparent from these and other theoretical concepts that, in a future quantum computer, the spins must be initialized, manipulated and read out one by one[3]. At about the same time, other researchers were independently developing 'toolkits' of sensitive spin-manipulation techniques to investigate fundamental quantum-mechanical processes in nanostructures such as decoherence on the atomic scale. Ultimately, around the start of this century, spintronics emerged[4], a field that seeks to encode classical information in the spin state of electrons. Both spintronics and quantum information processing have been major driving forces towards the control of single-spin systems.

Here we review experimental progress towards full control of the quantum states of single and coupled spins in different semiconductor systems. We also discuss the mechanisms that lead to the loss of spin coherence in these systems.

Single spins in semiconductors

Single-spin systems in semiconductors broadly fall into two categories: atomic impurities and quantum dots. Atomic impurities are routinely added to semiconductors to control the electrical properties (doping). When the concentration of impurities is very low, the possibility of addressing individual impurities arises. Atomic impurities may have nuclear spin, or they can act as a potential trap for electrons or holes. Often they do both, as in the case of phosphorus in silicon. If two or more impurities are present, or if there is a combination of impurities and lattice defects such as a vacancy, more complicated 'centres' can be formed that often have excellent properties for single-spin studies. One prime example is the nitrogen–vacancy (N–V) colour centre in diamond, which consists of a substitutional nitrogen atom next to a missing carbon (the vacancy) (Fig. 1). This N–V centre has a paramagnetic electron spin and a strong optical transition at a visible wavelength, which allows optical imaging of single spins.

Quantum dots, by contrast, behave like atoms in many ways, but they are fabricated in the laboratory. By engineering the electronic band structure, reducing the size of the semiconductor crystal in one or more dimensions, or applying electric fields, charge carriers can be confined to a small region of the crystal. If the region is roughly the same size as the wavelength of the charge carrier, the energy levels will be quantized as in real atoms. Many atomic properties, such as shell structure and optical selection rules, have analogues in quantum dots, giving rise to their nickname 'artificial atoms'[5–7]. In contrast to real atoms, however, quantum dots allow flexible control over the confinement potential and tend to be easier to excite optically. Quantum dots with large tunnel coupling (that is, strong overlap of their electronic wavefunctions) can form 'artificial molecules'. Such covalent bonding transforms the single-dot orbitals into molecular-like orbitals that span both quantum dots. As a consequence, spins in neighbouring coupled quantum dots overlap strongly and will form two-particle wavefunctions such as spin singlet and triplet states[8].

Quantum dots come in various sizes and in a range of materials. Here we mainly focus on the two types of quantum dot in which coherent dynamics have been observed at the single-spin level. In the first type, confinement is achieved through the application of electric fields, and measurements typically involve the transport of charge carriers through the device. Quantum dots with a tunable number of electrons are routinely fabricated from a two-dimensional electron gas (2DEG) that confines the charge carriers to a plane. Confinement in the remaining two dimensions is achieved by electric fields, either through metallic surface gates above the 2DEG (Fig. 1a) or, if a small pillar has been prepared by etching, from the edges. Gallium arsenide (GaAs) has been the material of choice for many years for these devices, as the high level of control has led to high-purity, flexible devices. More recently, motivated by the detrimental effect of lattice nuclear spins on the coherence times of electron spins, quantum dots have also been studied in materials such as silicon and carbon that can be isotopically purified to obtain a lattice that is free of nuclear spins.

[1]Kavli Institute of Nanoscience Delft, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands. [2]California Nanosystems Institute, University of California, Santa Barbara, California 93106, USA.

The second type of quantum dot is defined in the semiconductor during the growth of the crystal. For instance, small islands of semiconductor material such as indium gallium arsenide (InGaAs) can be created within a matrix of a semiconductor with a larger bandgap, such as GaAs (Fig. 1b). The difference in bandgap confines charge carriers to the island. Once the material is grown, the bandgap profile is fixed. However, changes to the overall potential, and potential gradients on top of the bandgap profile, can be induced by electric or magnetic fields. Another example of growth-defined dots is nanocrystal quantum dots, whose small size confines charge carriers. Double dots can be formed in nanocrystal dots by growing shells of different materials around the core.

Optical transitions in this second type of quantum dot typically have a large oscillator strength, and many studies use only optical techniques. Recent years have also seen the advent of hybrid systems, in which both electrical transport and optical excitation and detection are possible[9].

Experiments on single spins in quantum dots

In the 1990s, measurements of electron transport through single quantum dots yielded information about spin states[10]. The past five years have seen tremendous progress towards the control of single spins[8]. Single-spin dynamics was first studied in a series of pioneering experiments[11] at the NTT Basic Research Laboratories in Atsugi, Japan, in 2001 that made use of fast voltage pulses on gate electrodes. Toshimasa Fujisawa, Seigo Tarucha and co-workers found that if a transition between two states was forbidden by spin-selection rules, the corresponding decay time (more than 200 μs) was more than four orders of magnitude greater than for transitions not involving a change of spin (about 10 ns). In a second experiment, they made a single electron oscillate coherently between orbitals in neighbouring coupled dots[12]. The orbital ('charge') coherence of this oscillation was found to disappear in just a few nanoseconds, whereas theory was predicting coherence times of several microseconds for the spin degree of freedom[13–15].

In 2004, Leo Kouwenhoven and co-workers at the Kavli Institute of Nanoscience in Delft, the Netherlands, combined the pulse schemes of Fujisawa's group with a fast charge sensor that could tell exactly when an electron was entering or leaving the dot. By making the tunnelling rate of the electron from the dot dependent on its spin state, they could determine the spin state by measuring the charge on the dot over time (Fig. 2a). Two variations of this spin-to-charge conversion were demonstrated to work in single-shot mode[16,17]. Again, relaxation times for a single electron and for two-electron spin states were found to be of the order of a millisecond. A few years later, even longer electron spin relaxation times, of up to a second, were found at magnetic fields of a few tesla by Marc Kastner's group at the Massachusetts Institute of Technology in Cambridge[18].

Coherent control over two-electron spin states

Two electrons in neighbouring quantum dots with a significant tunnel coupling form a two-particle spin wavefunction, which can be a spin singlet or a spin triplet. The energy difference between these states can be described as an effective exchange splitting, $J(t)$. Control over this exchange splitting allows dynamical control of the two-electron spin states. If two electrons with opposite spin orientation in neighbouring dots are initially decoupled, turning on the coupling will result in a precession of the two spins in the singlet–triplet basis. This leads to periodic swapping of the two spin states at integer multiples of the time interval $\pi\hbar/J$ (where $\hbar$ is $h/2\pi$ and h is Planck's constant), whereas the electrons are entangled for intermediate times[1]. In fact, the state swapping occurs for arbitrary initial states of the two spins. This two-spin control, appropriately called a SWAP operation, is an essential ingredient for many proposals for quantum computing with spins in dots[19–21]. If logical quantum bits (qubits) are encoded in more than one spin, control over the exchange splitting is sufficient to build up any quantum gate[22]. The exchange operation has several benefits: the control is fully electrical, the interaction can be turned on and off, and the resultant gate operation times can be very short (less than a nanosecond).

The first step towards the exchange operation was the observation by Tarucha's group[23] of Pauli spin blockade in a double quantum dot. The presence of double-dot singlet and triplet states became apparent when the current was suppressed in one bias direction (Fig. 2c). It was later found that this current blockade can be lifted by fluctuating fields from the nuclear spins that cause mixing of the singlet and triplet spin states[24,25]. In 2005, by using the strength of the exchange interaction to control the mixing, Charles Marcus's group at Harvard University in Cambridge, Massachusetts, demonstrated coherent oscillations of two spins[26]. Although it was not yet possible to probe arbitrary input states, this experiment demonstrated the essence of the SWAP gate.

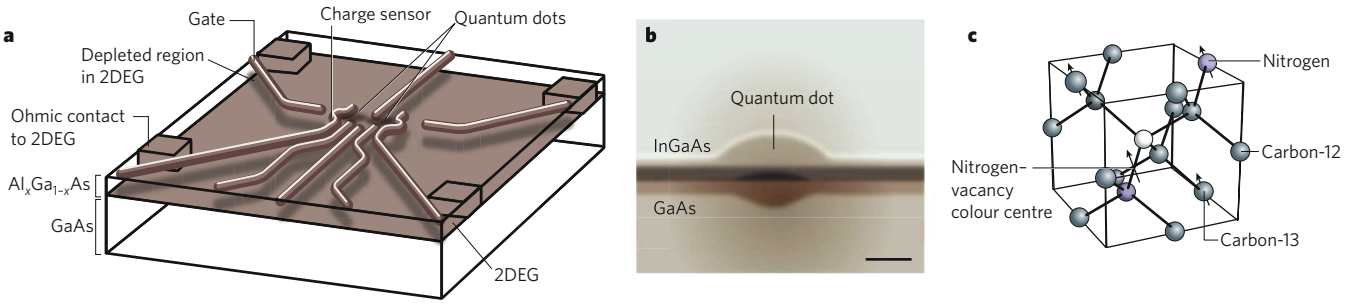


Figure 1 | Single-spin systems. Studies of the coherence of a single spin require a system in which the spin is localized and isolated from environmental disturbances. In semiconductors, such systems are either impurity atoms or quantum dots, which act as artificial atoms. In the three systems on which this article mainly focuses, the level of experimental control is so high that the dynamics of a single spin can be studied and manipulated. **a**, A quantum dot defined in a two-dimensional electron gas (2DEG). The electrons are confined in the third dimension by electric fields from the surface gate electrodes. Electron spins can be manipulated using magnetic resonance or a combination of electric fields and a position-dependent effective magnetic field. Interactions between spins in neighbouring tunnel-coupled dots are mediated by the exchange interaction. These quantum dots are typically measured at temperatures below 1 K. **b**, A quantum dot defined by growth. The semiconductor of the island has a smaller bandgap than that of the surrounding matrix, thereby confining charge carriers to the island. Spins can be created and controlled optically. Additional gates can be used to apply an electric field to the structure to change the number of carriers on the quantum dot. Measurements are typically carried out at around 4 K. Scale bar, 5 nm. **c**, A nitrogen–vacancy (N–V) colour centre in diamond, consisting of a substitutional nitrogen atom next to a missing carbon atom. The N–V centre (in the negatively charged state) comprises six electrons that form a spin triplet in the electronic ground state. Strong optical transitions to excited states, in combination with spin-selection rules, allow optical initialization and read-out of the electron spin. Coherent control of the spin has been demonstrated with high fidelity at room temperature using magnetic resonance. The N–V centre interacts with nearby electron spins by means of magnetic dipolar coupling, and through hyperfine interaction with nearby nuclear spins. Also, non-local coupling between N–V centres may be established by using the optical transition; photons then act as mediators of the interaction.

Single-spin rotations

A year after the coherent two-spin experiments, the Delft group, now headed by Lieven Vandersypen, demonstrated single-spin control[27] through magnetic resonance. In this technique, an oscillating magnetic field is applied perpendicular to the static magnetic field. When the frequency of the oscillating field is matched to the energy difference of the two spin states, the spins are rotated coherently.

Although electric fields do not couple directly to the spin, a coupling between the two can be mediated through a position-dependent effective magnetic field. By 'shaking' the electron in this field gradient, an oscillating effective magnetic field is imposed on the electron that can coherently rotate the spin (see, for example, ref. 28). A few examples of this approach have already been demonstrated in a quantum dot by exploiting a gradient in the nuclear spin polarization[29], a field gradient from a micromagnet[30], and the spin–orbit coupling[31]. In the last case, coherent control has been achieved on a timescale similar to that obtained with magnetic resonance (about 100 ns for a single rotation). In comparison to magnetic resonance, electrical control has the important advantage that it allows spins to be easily addressed locally, because electric fields are much easier to confine to small regions of space than magnetic fields.

Experiments on optically measured quantum dots

The physics of optically measured quantum dots is very similar to that of those studied electrically, but the experimental techniques differ markedly. Experiments on quantum dots in group III–V and group II–VI semiconductors, such as InGaAs dots in a GaAs matrix, make use of optical-selection rules in these materials. Shining circularly polarized light onto the material excites electron–hole pairs with specific spin. This has become a standard method for exciting packets of spin-polarized electrons in semiconductors and studying their coherent behaviour[32].

In a quantum dot, the same technique applies but with limited space for charge carriers. With proper tuning, the number of excited electron–hole pairs in the dot can be limited to one. In this way, a single electron and single hole can be created with well-defined spin states, in addition to any permanent charge carriers in the dot. The spin-selection rules also work the other way: when an electron–hole pair recombines, the polarization of the emitted photon tells us what the spins of the electron and the hole were. Optical-selection rules thereby allow the initialization and read-out of the spin states (Fig. 2d).

Optical techniques have been used to probe the stability of electron spins. In 2004, Jonathan Finley and co-workers at the Walter Schottky Institute in Munich, Germany, optically pumped electron–hole pairs that had a specific spin orientation into a large number of quantum dots. They then removed the holes by rapidly changing the electrical potential of the dots[33]. After a variable time, they reinserted a hole into each dot to allow recombination, and monitored the polarization of the emitted photons, which reflects the spin of the captured electrons. In these ensemble measurements, the electron spin could be found in the same orientation even after 20 ms. Finley and co-workers have recently repeated the spin relaxation measurements for single holes[34]. For a long time, it was thought that these hole spins would lose their orientation quickly as a result of strong spin–orbit coupling in the valence band. However, Finley's data pointed to very long hole-spin relaxation times of up to 300 μs, as predicted by a recent theory from Loss and co-workers at the University of Basel, Switzerland, that takes into account the confinement potential and strain[35]. Future experiments will seek to obtain coherent control of the hole spin state and determine the spin coherence time.

The spin orientation of electrons can also be inferred from the Kerr effect, in which the linear polarization of an incident laser beam is rotated in proportion to the spin polarization of electrons. This powerful technique has become a standard method for studying spin dynamics in semiconductors. It has recently been extended to the single-spin limit by the group of David Awschalom at the University of California, Santa Barbara[36], and subsequently by the group of Atac Imamoglu at ETH Zurich, Switzerland[37]. With this single-spin sensitivity, time-resolved observation of the precession of a single spin in a magnetic field has been achieved[38].

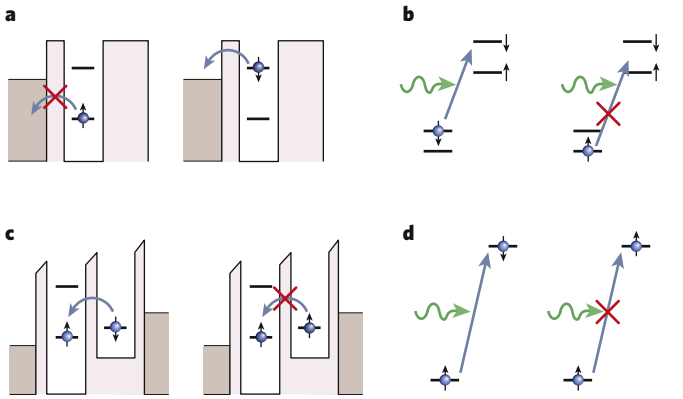


Figure 2 | Single-spin read-out. Studying a single spin is difficult because the magnetic moment of a spin is very small. Several spin read-out techniques have been developed in which the spin information is transferred to quantities that are more easily measured, such as electric charge or the polarization of light. This conversion requires that a transition between two states depends on the initial spin state; several examples of such transitions that are used in experiments are shown. **a**, **b**, Conversion of spin-state information into electric charge or photons by exploiting the energy difference between spin states. In **a**, an electron can tunnel from the quantum dot to the reservoir only if it is in the spin-down state. Measurement of the charge on the dot yields the spin state. **b**, A colour centre or quantum dot is optically excited and subsequently emits a photon only if it is in the spin-down state. The laser light is not resonant for the other spin state. Using a sensitive photon counter, the spin state can be determined after several optical cycles[65]. **c**, **d**, Spin read-out by spin-selection rules. The Pauli principle forbids two electrons with the same spin orientation to occupy a single orbital. Therefore, if one electron occupies an orbital, a second electron cannot enter if it has the same spin. Transitions that conserve spin (such as tunnelling and electric dipole transitions) can thus be blocked for certain spin states, hence the name 'Pauli spin blockade'. **c**, In a double quantum dot, the transition from the right dot to the left dot is blocked if the two electrons involved have the same spin. The second electron needs to go into a higher orbital, which is energetically not available. **d**, Circularly polarized laser light excites electrons with a certain spin orientation out of the valence band to the lowest orbital in the conduction band in a quantum dot. If an electron with the same spin orientation is already present in that orbital, the transition is forbidden.

Optical techniques also allow the coherent manipulation of spins. One method that has been proposed in the context of quantum information processing makes use of Raman transitions of spins in a microcavity[39]. Alternatively, single spins may be manipulated using the a.c. Stark effect[40], in which an intense laser pulse at a frequency slightly below the optical transition renormalizes the energy of the optical transition. When circularly polarized light is used, only one of the two spin states is affected by the laser pulse, resulting in an energy shift between spin up and spin down. This shift, known as the a.c. Stark shift, acts as an effective magnetic field along the light propagation direction; the magnitude of this field depends both on the detuning of the laser with respect to the optical transition and on the intensity of the pulse. Awschalom's group recently used the a.c. Stark effect to manipulate a single electron spin[41]. Short laser pulses were shown to induce rotations of the spin over an angle up to 180° in a time interval as short as 30 ps. This is about three orders of magnitude faster than any magnetic or electrical manipulation on single spins in quantum dots achieved thus far and is an important improvement in the context of quantum error correction.

Loss of spin coherence in quantum dots

In this discussion, we distinguish between energy relaxation processes (typically characterized by a spin relaxation time, T_1) and phase relaxation processes (characterized by a spin coherence time, T_2). By definition, T_1 sets a bound on T_2 such that $T_2 \leq 2T_1$. For successful quantum error correction, T_2 must exceed the spin manipulation time by several

orders of magnitude. A third timescale, T_2^*, is often used to denote the time after which the electron phase is randomized during free evolution. If the spin manipulation time is less than T_2^*, the fidelity of the control can be severely reduced, which adds a second requirement for quantum information application.

Quantum coherence of spins in semiconductor quantum dots is limited by coupling to other degrees of freedom in the environment. Electrons or holes can couple to states outside the quantum dot (Fig. 3a), and fluctuations in the electrical potential can indirectly lead to decoherence of the spin (Fig. 3b).

The absence of inversion symmetry in the lattice and the presence of electric fields or confinement asymmetries lead to coupling between spin and the motion of electrons (Fig. 3c). This spin–orbit coupling mixes the spin eigenstates. Except for small energy splitting, spin relaxation in group III–V quantum dots is typically dominated by spin–orbit coupling in combination with phonon emission that takes away the excess energy. Measurements of the spin relaxation time in many different devices have confirmed the theoretically predicted dependence on magnetic field and temperature[8]. However, the phase of localized electron spins is much less sensitive to the spin–orbit coupling[15]. The spin decoherence time, T_2, of electrons in group III–V quantum dots is typically limited by the nuclear spins (Fig. 3d).

The hyperfine interaction with the nuclear spins has two effects on the electron spin[42]. First, each nuclear spin exerts a tiny effective magnetic field on the electron spin. The sum of the fields of the roughly 1 million nuclear spins in a quantum dot, known as the Overhauser field, can be large (up to several tesla) if the nuclear spins all point in the same direction. The magnetic moment associated with the nuclear spins is small, so the thermal polarization is tiny even at millikelvin temperatures. However, the Overhauser field still fluctuates around this tiny average. A simple estimate tells us that for n nuclear spins, the statistical variation is of the order of $\sqrt{n}$, which corresponds to an effective magnetic field of a few millitesla for a typical group III–V quantum dot. Such a field causes the phase of the electron spin to change by π in roughly 10 ns. A measurement usually lasts tens of seconds, during which time the nuclear spins change orientation many times. One measurement therefore yields an average over many different nuclear spin configurations, leading to random phase variations between successive measurements. This leads to a dephasing time, T_2^*, of about 10 ns (refs 13, 14), a timescale that was first verified in optical experiments[43,44].

The Overhauser field changes slowly relative to the spin manipulation time, because the nuclear spins interact weakly both among themselves and with their surroundings. For example, recent optical experiments indicate that, in certain circumstances, nuclear spin polarizations in quantum dots can sometimes survive for up to an hour[45]. Simple spin-echo techniques can therefore be used to eliminate the effect of the quasi-static Overhauser field, provided that the electron spin can be manipulated on a timescale that is short compared with the spin precession time in the Overhauser field. There are two approaches to achieving this. The most straightforward is to make the manipulation time very short, either by using the exchange energy in two-spin systems or by optical manipulation using the a.c. Stark effect. Alternatively, the Overhauser field can be made smaller. One way of doing this is to narrow the distribution of the Overhauser fields by bringing the nuclear spins to a specific and stable quantum state[46–48]. Another option is to polarize all of the nuclear spins. Nuclear spin polarizations of up to 60% have been measured in quantum dots[44,49], but it is anticipated that a polarization far above 90% is required for a significant effect[50].

Another effect of the nuclear spins on the electron spin coherence comes from flip-flop processes[42], in which a flip of the electron spin (say from spin up to spin down) is accompanied by a flop of one nuclear spin (from spin down to spin up). In a first-order process, this leads to spin relaxation (the electron spin is flipped). If the electron spin is continuously repolarized, for example by optical pumping, the nuclear spins will all be flopped into the same spin state. After many such flip-flop events, a significant nuclear spin polarization can arise. This process is called dynamical nuclear polarization. If there is a large energy mismatch between the electron spin splitting and the nuclear spin splitting (because there is an external magnetic field, for instance), this first-order process is strongly suppressed. Second-order processes — in which two nuclear spins exchange their state by two flip-flops with the electron spin — are still possible. Through these virtual flip-flops, the nuclear spins can change orientation much faster than is possible with the magnetic dipolar interaction with nearby nuclear spins. This effectively leads to spin diffusion. The observed T_2 of about a microsecond is thought to be compatible with this picture, although firm experimental evidence isolating the different causes of nuclear field fluctuations is still lacking[8].

Spins of holes in the valence band of group III–V semiconductors have wavefunctions that have zero weight at the position of the nuclei, so the contact hyperfine interaction should not affect the coherence of holes. Richard Warburton and co-workers have recently initialized single hole spins in quantum dots at zero magnetic field[51] by adapting a procedure that was previously demonstrated on single electron spins[52].

The detrimental effect of the nuclear spins on the coherence in quantum dots has also spurred research into materials systems that contain

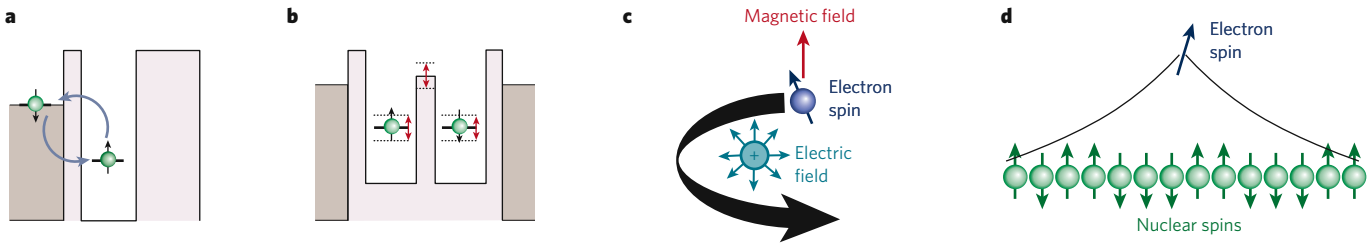


Figure 3 | Spin decoherence in quantum dots. The coherence of spins in quantum dots is affected by several mechanisms. **a**, Co-tunnelling. Although energy conservation forbids first-order tunnelling of charge carriers to states outside the dot at higher energy, second-order tunnelling processes (co-tunnelling) — in which a charge carrier tunnels from the dot to a reservoir and is replaced by a different charge carrier from the reservoir — are allowed[83]. The charge carrier from the reservoir will in general not be in the same spin quantum state as the one that first occupied the dot, so this process causes spin coherence to be lost. By increasing the energy difference between the dot and the reservoir states, and also making the tunnel coupling between them small, co-tunnelling processes can effectively be suppressed. **b**, Charge noise. Fluctuations in the electrical potential (charge noise) do not couple directly to the spin but can influence the spin dynamics indirectly. For example, the energy splitting, J, between singlet and triplet states in a double quantum dot depends strongly on the height of the tunnel barrier between the dots and the alignment of the levels in the dots. Any changes in the electrostatic environment can lead to changes (indicated by red arrows) in the barrier height and level misalignment, which modify J and therefore induce random phase shifts between the singlet and triplet states[84,85]. Charge switching and gate-voltage noise are two possible causes for such changes[86]. **c**, Spin–orbit coupling. The coupling between the spin and orbital of charge carriers leads to mixing of the spin states in a quantum dot. As a result of this coupling, any disturbance of the orbitals leads to phase fluctuations of the spin state. **d**, Nuclear spins. The charge carriers in the dot couple to the nuclear spins of the host material. These nuclear spins exert an effective magnetic field, and allow spin flip-flop processes that lead to spin relaxation and decoherence.

fewer or no nuclear spins. Two prominent examples are carbon and silicon, which can both be purified isotopically to yield a zero-spin lattice. Single and double quantum dots have been studied in these systems for several years, with control now approaching the level of GaAs systems[53–57]. Experiments probing the spin coherence times in silicon and carbon quantum dots are expected in the near future.

Coherent control of magnetic dopants

In contrast to non-magnetic nanostructures, individual magnetic-ion spins can be doped within group II–VI semiconductor quantum dots and measured through their exchange coupling to the electrons and holes. Statistically, it is possible to find an ion-impurity spin that is randomly doped at the centre of a single quantum dot. Using self-assembled quantum dots consisting of cadmium telluride and zinc telluride, Lucien Besombes *et al.*[58] isolated an individual paramagnetic manganese ion within individual dots. The micro-photoluminescence spectrum of an exciton (an electron–hole pair) was observed to split into six equally spaced lines owing to the quantization of manganese with a spin of 5/2. The next step was to apply a gate bias and change the charge state of the dot by pulling in either one electron or one hole. In this case, the coupling between the manganese ion and either the hole or the electron splits the six-line spectrum into twelve lines[59], in agreement with models based on spin exchange interactions within diluted magnetic semiconductors[60].

Dilute doping of group III–V semiconductors with manganese ions produces a unique environment for single-ion spin physics. Because the manganese states rest within the bandgap, it is not necessary to isolate a single manganese impurity within a single quantum dot, as the ions themselves act as recombination centres. In analogy with atomic physics, they form their own 'ideal' quantum dot states. The spin state of the manganese ion is independent of the electronic exciton and can be read out directly from the polarization of the manganese neutral acceptor emission[61]. In the absence of an applied magnetic field, the orientation of the magnetic ions is controlled by a dynamic interaction with optical injected electron spins. This mechanism is similar to dynamic nuclear polarization between electron spins and nuclear spins through the hyperfine interaction. After the manganese ions are partially aligned, a mean field interaction between the manganese ions, mediated by heavy hole states, favours a parallel alignment of the magnetic moments, creating a zero field splitting of the manganese-ion spins. The measurements indicate that single manganese-ion spins have longer coherence times than their electronic counterparts, motivating further studies of coherent control of manganese spins in semiconductors.

Coherent control of spins in diamond

Spins in diamond have recently become a leading candidate for solid-state quantum control, owing to their long coherence times and strong optical transitions, as well as the enormous progress that has been made in the growth and engineering of diamond as a unique semiconductor[62]. Ensemble experiments in the late 1990s indicated that spins of impurity centres in diamond can have very long coherence times, even at room temperature[63]. A more recent series of experiments demonstrated high-fidelity coherent control over electron spins and nuclear spins at room temperature at the single-spin level.

Most work is focused on the N–V centre (Fig. 1c) because of its attractive properties for quantum coherent operation[62]: the N–V centre's electronic-level structure allows both optical cooling and optical read-out of the electron spin. In 1997, following progress in confocal microscopy and the availability of diamond samples with a low concentration of N–V centres, Jörg Wrachtrup and co-workers reported the first study[64] of a single N–V-centre spin. In the seven years that followed, Fedor Jelezko, Wrachtrup and co-workers demonstrated single-shot read-out of the N–V electron spin at 1.5 K using resonant laser excitation[65] (Fig. 2b), coherent control of a single spin using magnetic resonance[66], and a two-qubit gate involving the host nuclear spin of the N–V centre[67]. In a parallel development, materials research achieved the growth of diamond using the chemical vapour deposition (CVD) method. With CVD, control over the number of impurity atoms increased enormously, resulting in the availability of very clean layers of diamond. Moreover, diamond nanocrystals of various sizes that contain N–V centres or other colour centres can now be grown[68]. These crystals have the advantage that they are small and light, and can be positioned onto other materials. This may facilitate interfacing spins in diamond with optical components such as fibres and cavities[69].

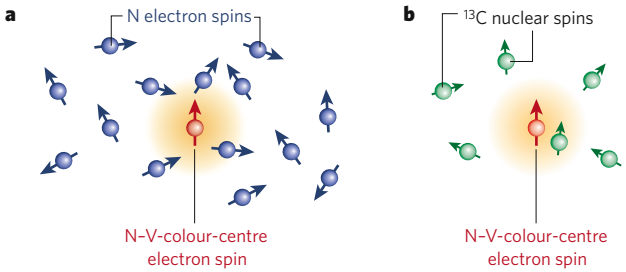


Figure 4 | Control and coherence in diamond. Spins in diamond are unique among solid-state systems in that single spins can be coherently controlled with high fidelity even at room temperature. The amount of impurity spins has a strong influence on coherence properties. **a**, Schematic representation of a nitrogen–vacancy (N–V) centre surrounded by electron spins of nitrogen impurities. In this case, the coherent dynamics of the N–V-centre spin are determined by the nitrogen atom's electron spins; the influence of nuclear spins is negligible because their magnetic moment is three orders of magnitude smaller than that of the electrons. Because an electron spin bath is easily tunable with a magnetic field, these systems allow detailed investigation of spin decoherence models and tests of quantum control in a tunable spin bath[73]. **b**, Schematic representation of an N–V centre surrounded by nuclear spins of carbon-13 in an ultrapure diamond. Nuclear spins that are much closer to the N–V centre than the others (within the orange sphere) stand out from the rest of the nuclear spins and can be individually distinguished and controlled[78].

The coherent evolution of impurity spins in diamond is dominated by magnetic interactions (Fig. 4). Whereas in quantum dots many nuclear spins have identical coupling to the electron spin, the highly localized nature of an impurity spin makes the magnetic interactions strongly dependent on the distance between spins. Other couplings such as a spin–orbit interaction have a much weaker effect at impurities than in quantum dots because of the much larger electronic-level splittings.

For impurity concentrations down to about 1 p.p.m., magnetic dipolar coupling between impurity electron spins dominates spin coherence in diamond[63,70] (Fig. 4a). In some cases, two spins are much closer to each other than to the rest of the spins, in which case the dynamics become a simple two-spin evolution. One example is an N–V centre in close proximity to a single nitrogen impurity, in which case the nitrogen spin can be polarized and read out through the N–V centre[71,72].

In general, an N–V centre will be coupled to many nitrogen spins, which can be viewed as a 'spin bath'. Even in this case, the spin of the N–V centre can be controlled with high fidelity, allowing investigation of decoherence induced by the spin bath. Analogous to an electron spin in a quantum dot that is coupled to a nuclear spin bath, the nitrogen electron spins influence the evolution of the N–V centre in two ways. First, the magnetic dipolar field from the bath shifts the energy splitting between the N–V centre's spin states (analogous to the Overhauser field in quantum dots). It has recently been shown[73] that the interactions within the bath, leading to the fluctuations of the dipolar field, are strongly suppressed when a magnetic field is applied, in which case single-spin flips do not conserve energy. The second effect of the spin bath comes from flip-flop processes with the N–V electron spin. In contrast to a nuclear spin bath (where the spin splitting is tiny), the electron spin bath can be tuned into energy resonance with the N–V-centre electron spin. Resonant flip-flop processes then provide a strong additional decoherence path, leading to much shorter spin coherence times.

Because the spin coherence in diamond is dominated by magnetic interactions, it is not strongly temperature dependent, except at low

temperatures and high magnetic fields, at which the bath spins polarize thermally; this occurs for electron spins in a field of 8 T at a few kelvin. Recent experiments on a diamond with a high concentration of nitrogen electron spins show that when all of the electron spins are thermally polarized, the fluctuations in the spin bath are completely frozen out and the spin coherence time reaches the same high value as in ultrapure diamond[74].

In most diamonds studied thus far, the positions of the impurity spins were random, affected only by growth parameters. In 2005, single N–V centres were deliberately created by ion-implanting nitrogen[75,76]. This approach may lead to fundamental studies of spin coherence in diamond by designing different spin environments, as well as allow pathways for engineering spin qubits into future scalable quantum information-processing systems.

In diamonds where the impurity concentration is very low (below 1 p.p.m.), the presence of the few nuclear spins of the carbon-13 isotope (which has a natural abundance of 1.1%) becomes apparent. These nuclear spins also constitute a spin bath, which limits the N–V centre's coherence time to a few hundred microseconds[70,71]. Because a single N–V spin can be rotated using magnetic resonance in less than 10 ns, more than 10,000 error-free operations can be performed, which is within the commonly assumed threshold for quantum error correction. As in the case of impurity electron spins, if a few nuclear spins are much closer to the N–V centre than the other nuclear spins, their individual coupling to the N–V-centre spin can be detected[77] (Fig. 4b). Mikhail Lukin's group at Harvard University demonstrated that these nuclear spins can be used to store quantum information for much longer than the electron spin's coherence time[78]. The quantum state of the N–V electron spin can be mapped onto, or retrieved from, the nuclear-spin memory through a combination of state-dependent precession of the nuclear spin and fast optical reinitialization of the N–V-centre spin. Experiments have shown that even on a 20-ms timescale, the nuclear spin shows no sign of decoherence[78], suggesting that nuclear spins may have coherence times of seconds or even longer. By extending the control to multiple nuclear spins, a small quantum memory can be created that will operate at room temperature.

As well as long spin coherence times, N–V centres also have a strong optical transition. Lifetime-limited optical linewidths have been observed[79], and the optical preparation of a coherent superposition of spin states has been demonstrated in coherent population trapping experiments on single N–V centres[80]. These results may, in the future, be extended to dynamical all-optical control of single spins in diamond.

Optical control opens the door to schemes for creating entangled states of spins at large distances[81], in a similar way as was recently demonstrated for atom traps[82]. Such long-distance entanglement is also a crucial ingredient for applications in quantum communication.

Outlook

After enormous progress in recent years, researchers can now initialize, control and read out single spins in semiconductors in a few specific systems, with others likely to be added to the list within a few years. The coherence times of electron spins in materials with few or no nuclear spins, as well as the coherence times of hole spins, are expected to be much longer than for electron spins in group III–V semiconductors. Carbon-based materials, such as carbon nanotubes and graphene, are being heavily investigated; diamond has already shown its potential for quantum coherence studies (at room temperature) with a level of single-spin control that meets the quantum information-processing error-correction threshold.

The emphasis of this research area will shift in the coming years from single-spin control to the creation and manipulation of entangled states of two or more spins, as well as the development of sophisticated quantum control techniques. This will lead the way for more studies on fundamental issues such as decoherence and the role of measurements in quantum mechanics. At the same time, protocols for quantum information processing may be tested in systems with few spins. These are exciting times for 'spin doctors', as they continue to drive a rapidly expanding field that has a promising future. ■

1. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57,** 120–126 (1998).
2. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393,** 133–137 (1998).
3. DiVincenzo, D. P. The physical implementation of quantum computation. *Fortschr. Phys.* **48,** 771–783 (2000).
4. Wolf, S. A. *et al.* Spintronics: a spin-based electronics vision for the future. *Science* **294,** 1488–1495 (2001).
5. Kastner, M. A. Artificial atoms. *Phys. Today* **46,** 24–31 (1993).
6. Ashoori, R. C. Electrons in artificial atoms. *Nature* **379,** 413–419 (1996).
7. Kouwenhoven, L. P. & Marcus, C. M. Quantum dots. *Phys. World* **11,** 35–39 (1998).
8. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79,** 1217–1265 (2007).
9. Zrenner, A. *et al.* Coherent properties of a two-level system based on a quantum-dot photodiode. *Nature* **418,** 612–614 (2002).
10. Kouwenhoven, L. P., Austing, D. G. & Tarucha, S. Few-electron quantum dots. *Rep. Prog. Phys.* **64,** 701–736 (2001).
11. Fujisawa, T., Austing, D. G., Tokura, Y., Hirayama, Y. & Tarucha, S. Allowed and forbidden transitions in artificial hydrogen and helium atoms. *Nature* **419,** 278–281 (2002).
12. Hayashi, T., Fujisawa, T., Cheong, H. D., Jeong, Y. H. & Hirayama, Y. Coherent manipulation of electronic states in a double quantum dot. *Phys. Rev. Lett.* **91,** 226804 (2003).
13. Khaetskii, A. V., Loss, D. & Glazman, L. Electron spin decoherence in quantum dots due to interaction with nuclei. *Phys. Rev. Lett.* **88,** 186802 (2002).
14. Merkulov, I. A., Efros, A. L. & Rosen, J. Electron spin relaxation by nuclei in semiconductor quantum dots. *Phys. Rev. B* **65,** 205309 (2002).
15. Golovach, V. N., Khaetskii, A. & Loss, D. Phonon-induced decay of the electron spin in quantum dots. *Phys. Rev. Lett.* **93,** 016601 (2004).
16. Elzerman, J. M. *et al.* Single-shot read-out of an individual electron spin in a quantum dot. *Nature* **430,** 431–435 (2004).
17. Hanson, R. *et al.* Single-shot readout of electron spin states in a quantum dot using spin-dependent tunnel rates. *Phys. Rev. Lett.* **94,** 196802 (2005).
18. Amasha, S. *et al.* Electrical control of spin relaxation in a quantum dot. *Phys. Rev. Lett.* **100,** 046803-1-4 (2008).
19. Levy, J. Universal quantum computation with spin-1/2 pairs and Heisenberg exchange. *Phys. Rev. Lett.* **89,** 147902 (2002).
20. Taylor, J. M. *et al.* Fault-tolerant architecture for quantum computation using electrically controlled semiconductor spins. *Nature Phys.* **1,** 177–183 (2005).
21. Hanson, R. & Burkard, G. Universal set of quantum gates for double-dot spin qubits with fixed interdot coupling. *Phys. Rev. Lett.* **98,** 050502 (2007).
22. DiVincenzo, D. P., Bacon, D.P., Kempe, J., Burkard, G. & Whaley, K. B. Universal quantum computation with the exchange interaction. *Nature* **408,** 339–342 (2000).
23. Ono, K., Austing, D. G., Tokura, Y. & Tarucha, S. Current rectification by Pauli exclusion in a weakly coupled double quantum dot system. *Science* **297,** 1313–1317 (2002).
24. Johnson, A. C. *et al.* Triplet–singlet spin relaxation via nuclei in a double quantum dot. *Nature* **435,** 925–928 (2005).
25. Koppens, F. *et al.* Control and detection of singlet–triplet mixing in a random nuclear field. *Science* **309,** 1346–1350 (2005).
26. Petta, J. R. *et al.* Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309,** 2180–2184 (2005).
27. Koppens, F. H. L. *et al.* Driven coherent oscillations of a single electron spin in a quantum dot. *Nature* **442,** 766–771 (2006).
28. Kato, Y. *et al.* Gigahertz electron spin manipulation using voltage controlled g-tensor modulation. *Science* **299,** 1201–1204 (2003).
29. Laird, E. A. *et al.* Hyperfine-mediated gate-driven electron spin resonance. *Phys. Rev. Lett.* **99,** 246601 (2007).
30. Pioro-Ladriere, M. *et al.* Electrically driven single-electron spin resonance in a slanting Zeeman field. Preprint at <http://arxiv.org/abs/0805.1083> (2008).
31. Nowack, K. C., Koppens, F. H. L., Nazarov, Y. V. & Vandersypen, L. M. K. Coherent control of a single electron spin with electric fields. *Science* **318,** 1430–1433 (2007).
32. Awschalom, D. D. & Kikkawa, J. M. Electron spin and optical coherence in semiconductors. *Phys. Today* **52,** 33–38 (1999).
33. Kroutvar, M. *et al.* Optically programmable electron spin memory using semiconductor quantum dots. *Nature* **432,** 81–84 (2004).
34. Heiss, D. *et al.* Observation of extremely slow hole spin relaxation in self-assembled quantum dots. *Phys. Rev. B* **76,** 241306 (2007).
35. Bulaev, D. V. & Loss, D. Spin relaxation and decoherence of holes in quantum dots. *Phys. Rev. Lett.* **95,** 076805 (2005).
36. Berezovsky, J. *et al.* Nondestructive optical measurements of a single electron spin in a quantum dot. *Science* **314,** 1916–1920 (2006).
37. Atature, M., Dreiser, J., Badolato, A. & Imamoglu, A. Observation of Faraday rotation from a single confined spin. *Nature Phys.* **3,** 101–106 (2007).
38. Mikkelsen, M. H., Berezovsky, J., Stoltz, N. G., Coldren, L. A. & Awschalom, D. D. Optically detected coherent spin dynamics of a single electron in a quantum dot. *Nature Phys.* **3,** 770–773 (2007).
39. Imamoglu, A. *et al.* Quantum information processing using quantum dot spins and cavity QED. *Phys. Rev. Lett.* **83,** 4204–4207 (1999).
40. Cohen-Tannoudji, C. & Dupont-Roc, J. Experimental study of Zeeman light shifts in weak magnetic fields. *Phys. Rev. A* **5,** 968–984 (1972).
41. Berezovsky, J., Mikkelsen, M. H., Stoltz, N. G., Coldren, L. A. & Awschalom, D. D. Picosecond coherent optical manipulation of a single electron spin in a quantum dot. *Science* **320,** 349–352 (2008).
42. Meier, F. & Zakharchenya, B. P. (eds) *Optical Orientation* (North-Holland, Amsterdam, 1984).
43. Braun, P. F. *et al.* Direct observation of the electron spin relaxation induced by nuclei in quantum dots. *Phys. Rev. Lett.* **94,** 116601 (2005).
44. Bracker, A. S. *et al.* Optical pumping of the electronic and nuclear spin of single charge-tunable quantum dots. *Phys. Rev. Lett.* **94,** 047402 (2005).
45. Giedke, G., Taylor, J. M., D'Alessandro, D., Lukin, M. D. & Imamoglu, A. Quantum measurement of a mesoscopic spin ensemble. *Phys. Rev. A* **74,** 032316 (2006).

46. Greilich, A. *et al.* Nuclei-induced frequency focusing of electron spin coherence. *Science* **317,** 1896–1899 (2007).
47. Stepanenko, D., Burkard, G., Giedke, G. & Imamoglu, A. Enhancement of electron spin coherence by optical preparation of nuclear spins. *Phys. Rev. Lett.* **96,** 136401 (2006).
48. Klauser, D., Coish, W. A. & Loss, D. Nuclear spin state narrowing via gate-controlled Rabi oscillations in a double quantum dot. *Phys. Rev. B* **73,** 205302 (2006).
49. Baugh, J., Kitamura, Y., Ono, K. & Tarucha, S. Large nuclear Overhauser fields detected in vertically coupled double quantum dots. *Phys. Rev. Lett.* **99,** 096804 (2007).
50. Coish, W. A. & Loss, D. Hyperfine interaction in a quantum dot: Non-Markovian electron spin dynamics. *Phys. Rev. B* **70,** 195340 (2004).
51. Gerardot, B. D. *et al.* Optical pumping of a single hole spin in a quantum dot. *Nature* **451,** 441–444 (2008).
52. Atatüre, M. *et al.* Quantum-dot spin-state preparation with near-unity fidelity. *Science* **312,** 551–553 (2006).
53. Mason, N., Biercuk, M. J. & Marcus, C. M. Local gate control of a carbon nanotube double quantum dot. *Science* **303,** 655–658 (2004).
54. Sapmaz, S., Meyer, C., Beliczynski, P. M., Jarillo-Herrero, P. D. & Kouwenhoven,L. P. Excited state spectroscopy in carbon nanotube double quantum dots. *Nano Lett.* **6,** 1350–1355 (2006).
55. Hu, Y. *et al.* Double quantum dot with integrated charge sensor based on Ge/Si heterostructure nanowires. *Nature Nanotech.* **2,** 622–625 (2007).
56. Simmons, C. B. *et al.* Single-electron quantum dot in Si/SiGe with integrated charge sensing. *Appl. Phys. Lett.* **91,** 213103 (2007).
57. Liu, H. W. *et al.* Pauli-spin-blockade transport through a silicon double quantum dot. *Phys. Rev. B* **77,** 073310 (2008).
58. Besombes, L. *et al.* Probing the spin state of a single magnetic ion in an individual quantum dot. *Phys. Rev. Lett.* **93,** 207403 (2004).
59. Léger, Y., Besombes, L., Fernández-Rossier, J., Maingault, L. & Mariette, H. Electrical control of a single Mn atom in a quantum dot. *Phys. Rev. Lett.* **97,** 107401 (2006).
60. Erwin, S. C. Nanomagnetism: spin doctors play with single electrons. *Nature Nanotech.* **1,** 98–99 (2006).
61. Myers, R. C. *et al.* Zero-field optical manipulation of magnetic ions in semiconductors. *Nature Mater.* **7,** 203–208 (2008).
62. Awschalom, D. D., Epstein, R. & Hanson, R. The diamond age of spintronics. *Sci. Am.* **297,** 84–91 (2007).
63. Reynhardt, E. C., High, G. L. & vanWyk, J. A. Temperature dependence of spin–spin and spin–lattice relaxation times of paramagnetic nitrogen defects in diamond. *J. Chem. Phys.* **109,** 84718477 (1998).
64. Gruber, A. *et al.* Scanning confocal optical microscopy and magnetic resonance on single defect centers. *Science* **276,** 2012–2014 (1997).
65. Jelezko, F., Popa, I., Gruber, A. & Wrachtrup, J. Single spin states in a defect center resolved by optical spectroscopy. *Appl. Phys. Lett.* **81,** 2160–2162 (2002).
66. Jelezko, F., Gaebel, T., Popa, I., Gruber, A. & Wrachtrup, J. Observation of coherent oscillations in a single electron spin. *Phys. Rev. Lett.* **92,** 76401 (2004).
67. Jelezko, F. *et al.* Observation of coherent oscillation of a single nuclear spin and realization of a two-qubit conditional quantum gate. *Phys. Rev. Lett.* **93,** 130501 (2004).
68. Rabeau, J. R. *et al.* Single nitrogen vacancy centers in chemical vapor deposited diamond nanocrystals. *Nano Lett.* **7,** 3433–3437 (2007).
69. Park, Y.-S., Cook, A. K. & Wang, H. Cavity QED with Diamond nanocrystals and silica microspheres. *Nano Lett.* **6,** 2075–2079 (2006).
70. Kennedy, T. A. *et al.* Long coherence times at 300 K for nitrogen–vacancy center spins in diamond grown by chemical vapor deposition. *Appl. Phys. Lett.* **83,** 4190–4192 (2003).
71. Gaebel, T. *et al.* Room-temperature coherent coupling of single spins in diamond. *Nature Phys.* **2,** 408–413 (2006).
72. Hanson, R., Mendoza, F. M., Epstein, R. J. & Awschalom, D. D. Polarization and readout of coupled single spins in diamond. *Phys. Rev. Lett.* **97,** 087601 (2006).
73. Hanson, R., Dobrovitski, V. V., Feiguin, A. E., Gywat, O. & Awschalom, D. D. Coherent dynamics of a single spin interacting with an adjustable spin bath. *Science* **320,** 352–355 (2008).
74. Takahashi, S., Hanson, R., van Tol, J., Sherwin, M. S. & Awschalom, D. D. Quenching spin decoherence in diamond through spin bath polarization. Preprint at <http://arxiv.org/abs/0804.1537> (2008).
75. Meijer, J. *et al.* Generation of single colour centers by focussed nitrogen implantation. *Appl. Phys. Lett.* **87,** 261909 (2005).
76. Rabeau, J. R. *et al.* Implantation of labelled single nitrogen vacancy centers in diamond using 15N. *Appl. Phys. Lett.* **88,** 023113 (2006).
77. Childress, L. *et al.* Coherent dynamics of coupled electron and nuclear spin qubits in diamond. *Science* **314,** 281–285 (2006).
78. Dutt, M. V. G. *et al.* Quantum register based on individual electronic and nuclear spin qubits in diamond. *Science* **316,**1312–1316 (2007).
79. Tamarat, P. *et al.* Stark shift control of single optical centers in diamond. *Phys. Rev. Lett.* **97,** 083002 (2006).
80. Santori, C. *et al.* Coherent population trapping of single spins in diamond under optical excitation. *Phys. Rev. Lett.* **97,** 247401 (2006).
81. Barrett, S. D. & Kok, P. Efficient high-fidelity quantum computation using matter qubits and linear optics. *Phys. Rev. A* **71,** 060310 (2005).
82. Moehring, D. L. *et al.* Entanglement of single-atom quantum bits at a distance. *Nature* **449,** 68–72 (2007).
83. Averin, D. V. & Nazarov, Y. V. Virtual electron diffusion during quantum tunnelling of the electric charge. *Phys. Rev. Lett.* **65,** 2446–2449 (1990).
84. Coish, W. A. & Loss, D. Singlet–triplet decoherence due to nuclear spins in a double quantum dot. *Phys. Rev. B* **72,** 25337 (2005).
85. Hu, X. & Das Sarma, S. Charge-fluctuation-induced dephasing of exchange-coupled spin qubits. *Phys. Rev. Lett.* **96,** 100501 (2006).
86. Jung, S. W., Fujisawa, T., Hirayama, Y. & Jeong, Y. H. Background charge fluctuation in a GaAs quantum dot device. *Appl. Phys. Lett.* **85,** 768–770 (2004).

Acknowledgements We thank the Air Force Office of Scientific Research (AFOSR), the Dutch Organization for Fundamental Research on Matter (FOM) and the Netherlands Organization for Scientific Research (NWO) for support.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to the authors (r.hanson@tudelft.nl; awsch@physics.ucsb.edu).

REVIEWS

Induction and effector functions of T_H17 cells

Estelle Bettelli[1], Thomas Korn[1]†, Mohamed Oukka[1] & Vijay K. Kuchroo[1]

T helper (T_H) cells constitute an important arm of the adaptive immune system because they coordinate defence against specific pathogens, and their unique cytokines and effector functions mediate different types of tissue inflammation. The recently discovered T_H17 cells, the third subset of effector T helper cells, have been the subject of intense research aimed at understanding their role in immunity and disease. Here we review emerging data suggesting that T_H17 cells have an important role in host defence against specific pathogens and are potent inducers of autoimmunity and tissue inflammation. In addition, the differentiation factors responsible for their generation have revealed an interesting reciprocal relationship with regulatory T (T_{reg}) cells, which prevent tissue inflammation and mediate self-tolerance.

The hallmark of adaptive immunity in advanced vertebrates is the existence of lymphocytes, which induce and regulate immune responses. When activated by pathogens in a specific cytokine environment, naive $CD4^+$ T cells differentiate into different subsets with distinct effector functions aimed at orchestrating and mobilizing other cell types to effectively clear invading pathogens. Based on cytokine phenotypes, initially the existence of two distinct effector T_H subsets was proposed: T_H1 and T_H2 (ref. 1). T_H1 cells produce interferon-γ (IFN-γ) and mediate protection against intracellular pathogens, whereas T_H2 cells produce interleukin-4 (IL-4), IL-13 and IL-25 (also known as IL-17E) and orchestrate the clearance of extracellular pathogens[1,2] (Fig. 1). Recently this paradigm has been updated following the discovery of a third subset of T_H cells; these cells, known as T_H17 cells (ref. 3), produce IL-17 and exhibit distinct effector functions. In the past four years there has been an explosion of information regarding this T-cell subset: the cytokines for their differentiation have been identified, the key transcription factors that are involved in their generation have been recognized and their function in tissue inflammation has been established. This review summarizes this information to develop a comprehensive view of the generation and function of T_H17 cells.

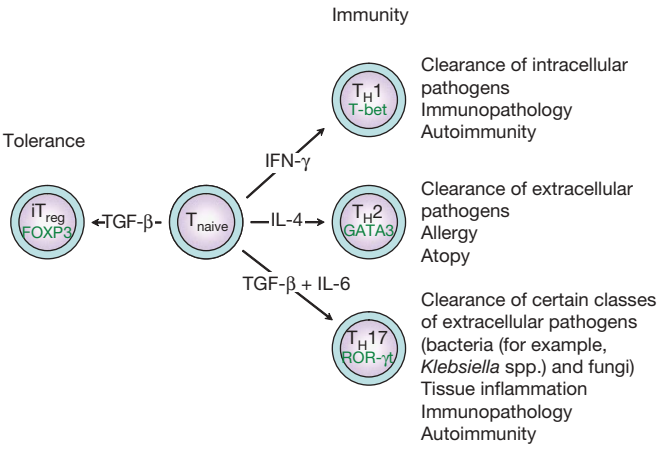


Figure 1 | Subsets of T helper cells. Depending on the cytokine milieu present at the time of the initial engagement of their T-cell receptor and co-stimulatory receptors in the peripheral immune compartment, naive $CD4^+$ T cells can differentiate into various subsets of T helper cells (T_H1, T_H2 and T_H17). However, in the presence of TGF-β, naive T cells convert into FOXP3-expressing induced T_{reg} (iT_{reg}) cells. For each T helper cell differentiation programme, specific transcription factors have been identified as master regulators (T-bet, GATA3 and ROR-γt). Terminally differentiated T helper cells are characterized by a specific combination of effector cytokines that orchestrate specific and distinct effector functions of the adaptive immune system.

T_H17 cells and T_H17-specific effector cytokines

T_H17 cells are characterized by the production of IL-17A (also called IL-17), IL-17F and IL-22 (Box 1) and are thought to clear extracellular pathogens not effectively handled by either T_H1 or T_H2 cells (Fig. 1). Because T_H17 cells produce large quantities of IL-17A, most T_H17-mediated effects are attributed to this cytokine. IL-17A is the prototypic cytokine of the IL-17 family, which includes six members: IL-17A, B, C, D, E and F[4]. IL-17 is a phylogenetically old cytokine that is also detected in non-mammalian vertebrates[5].

In addition to IL-17A, T_H17 cells co-produce IL-17F[3,6]. IL-17A and IL-17F have similar functions. They induce the production of proinflammatory cytokines, chemokines and metalloproteinases from various tissues and cell types (Box 1). As a result, they recruit neutrophils to tissues.

Although there is often a coordinated expression of IL-17A and IL-17F in T_H17 cells and other cell types, it is now clear that there are T_H cells expressing only IL-17A, IL-17F or an IL-17A–IL-17F heterodimer[7,8]. In addition to IL-17A and IL-17F, T_H17 cells produce other effector cytokines, namely IL-21 and IL-22 (refs 6, 9–12). Neither IL-21 nor IL-22 are T_H17-exclusive cytokines, but are preferentially expressed in T_H17 cells.

Recent work from our group and others showed that IL-21, a member of the IL-2 family of cytokines, is produced in large amounts by T_H17 cells and $ICOS^+CXCR5^+CCR7^+$ T follicular helper cells[13]. These T follicular helper cells home to the B-cell areas of secondary lymphoid tissue and provide cognate help to B cells. However, it remains to be seen whether T follicular helper cells could represent activated or memory T_H17 cells that help B cells in secondary or tertiary lymphoid organs.

[1]Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. †Present address: Neurologische Klinik, Klinikum rechts der Isar, Technische Universität Munich, Germany.

Box 1 | Effector cytokines produced by T_H17 cells

IL-17A

Source. T_H17 cells[3,55,56], $CD8^+$ T cells[73], γδ T cells[74], neutrophils[75], eosinophils[76] and monocytes[77].

Receptor(s). IL-17RA is the cognate receptor for IL-17A. It is expressed at high levels on haematopoietic cells, and at lower levels on osteoblasts, fibroblasts, endothelial cells and epithelial cells[78]. Human IL-17RC binds human IL-17A with high affinity, but mouse IL-17RC does not bind mouse IL-17A[79]. Human IL-17RA–IL-17RC can form a heterodimer that binds human IL-17A[80].

Effects. IL-17A induces pro-inflammatory cytokines (IL-6, TNF-α and IL-1β[78]) and chemokines (CXCL1, GCP-2, CXCL8 or IL-8, CINC, MCP-1; ref. 81). It increases the production of prostaglandin E2 (ref. 81), nitric oxide[82] and matrix-metalloproteinases[55], increases the recruitment of neutrophils[14,75] and modulates neutrophil homeostasis[83].

IL-17F

Source. T_H17 cells[3,6,9], monocytes[84] and possibly other cell types.

Receptor(s). IL-17RC is the cognate receptor for IL-17F. It is expressed at low levels on haematopoietic cells, and at high levels on non-haematopoietic cells[80]. Human IL-17RA–IL-17RC heterodimers can bind human IL-17F[80].

Effects. IL-17F induces pro-inflammatory cytokines (IL-6, ref. 85) and chemokines (CXCL1, GCP-2, CXCL8 or IL-8, ref. 85), and increases the recruitment of neutrophils[86].

IL-22

Source. T_H17 cells[6,9], activated T cells and natural killer cells[87].

Receptor(s). The receptor for IL-22 is a heterodimer consisting of IL-22R1 and IL-10R2 (ref. 88). IL-10R2 is ubiquitously expressed in haematopoietic and non-haematopoietic cells[89]. IL-22R1 (CRF2-9) is expressed on a variety of epithelial and parenchymal tissues (skin, liver, kidney, pancreas, intestine and lung)[90].

Effects. IL-22 increases acute-phase reactants in hepatocytes[91] and protects them from acute liver inflammation[35]. It induces the expression of β-defensins in epithelial cells[6,90] and promotes epidermal hyperplasia[9].

IL-21

Source. $CD4^+$ T cells stimulated with IL-6, T_H17 cells[10–12], T follicular helper cells[13], natural killer cells and natural killer T cells[10,92].

Receptor(s). The receptor for IL-21 is a heterodimer consisting of common cytokine-receptor γ chain (γ_c) and IL-21R[93]. γ_c is expressed in lymphoid, but not in non-lymphoid and non-haematopoietic cells[94]. IL-21R is restricted to haematopoietic cells with highest levels of expression on B cells, but also on T cells, natural killer cells, and some populations of myeloid cells[92].

Effects. IL-21 participates in the differentiation/amplification of T_H17 cells[10–12]. In combination with IL-7 or IL-15, IL-21 stimulates the proliferation and differentiation of $CD8^+$ T cells[95,96]. It promotes B-cell differentiation and antibody class switching (IgG1, IgG3)[96], induces the differentiation and cytotoxic programme of natural killer cells[92] and natural killer T cells[97], and induces CXCL8 in macrophages[98].

IL-22 is a member of the IL-10 family of cytokines, produced by activated T cells and natural killer cells. It mediates its effects through a receptor complex composed of the IL-10R2 and the IL-22R chains. Interestingly, high concentrations of transforming growth factor-β (TGF-β) can inhibit IL-6-induced IL-22 expression[9]. Furthermore, whereas the combination of TGF-β plus IL-6 induces large quantities of IL-17A and IL-17F by T_H17 cells, the secretion of large amounts of IL-22 by T_H17 cells requires the addition of IL-23 *in vitro*[6,9]. This suggests that IL-22 could represent an end point effector cytokine produced by terminally differentiated T_H17 cells.

An important role for T_H17 cells in host defence

So far it is unclear which class of pathogens preferentially induces a T_H17 response because pathogens as diverse as the Gram-positive *Propionibacterium acnes*, the Gram-negative *Citrobacter rodentium*, *Klebsiella pneumoniae*, *Bacteroides* spp. and *Borrelia* spp., the acid-fast *Mycobacterium tuberculosis*, and fungi such as *Candida albicans* can all trigger a substantial T_H17 response[14–19]. The fungal-cell-wall-derived product zymosan or other dectin-1 ligands as well as muramyldipeptide (a derivative of bacterial peptidoglycan) are able to promote IL-17 production in T cells[20]. Therefore, T_H17 responses are likely to emerge as an early response to a number of pathogens not handled well by T_H1- or T_H2-type immunity and which require robust tissue inflammation to be cleared.

Indeed, T_H17 cells appear at sites of inflammation with rapid kinetics. Through the potent induction of chemokines, T_H17 cells could bridge the gap between innate and adaptive immunity and attract other subsets of T helper cells to sites of infection at later stages of the inflammatory process. This has most convincingly been shown for *M. tuberculosis* infection, for which an early T_H17 response is required to bring T_H1 cells into the infected lung tissue to control the infection[18].

T_H17 cells in autoimmune diseases

It is widely conceived that organ-specific autoimmune diseases are the result of dysregulated autoantigen-specific T_H1 responses. In many animal models of human autoimmune diseases, T_H1 cells have been shown to be pathogenic[21]. However, the concept that autoimmune diseases were exclusively mediated by T_H1 cells has been challenged, and the idea that T_H17 cells are an important part of the autoimmune reaction has emerged in light of the following observations: first, mice deficient for the T_H1 effector cytokine IFN-γ develop enhanced experimental autoimmune encephalomyelitis (EAE)[22]; second, deficiency in the IL-12p35 subunit (specific for IL-12) does not alter the progression of EAE, but deficiency in either p40 or p19, which form IL-23, results in a decreased number of T_H17 cells and protection from EAE and collagen-induced arthritis[23,24]; third, the transfer of myelin-reactive IL-17-producing T cells expanded with IL-23 *in vitro* induces severe EAE[3]; and fourth, IL-17 has profound pro-inflammatory effects and induces tissue damage during the course of various autoimmune diseases. Indeed, IL-17 can directly or indirectly promote cartilage and bone destruction[25]. Conversely, IL-17-deficient mice develop attenuated collagen-induced arthritis[26] and EAE[27]. Increased levels of IL-17 have been observed in patients with rheumatoid arthritis[28], multiple sclerosis[29], inflammatory bowel disease[30] and psoriasis[31]. Furthermore, IL-22 produced by T_H17 cells mediates IL-23-induced acanthosis and dermal inflammation[9]. In addition, IL-22, similarly to IL-17, can disrupt tight junctions between endothelial cells of the blood–brain barrier[32]. These data indicate a pathogenic role of T_H17-associated cytokines and T_H17 cells in inducing autoimmune tissue inflammation both in experimental animals and in humans.

Despite the recent major interest in T_H17 cells, these cells may not be the only T_H cells that can induce autoimmunity because T_H1 cells can readily transfer organ-specific autoimmune disease[33]. It is therefore possible that there is a sequential involvement and different functions of T_H17 and T_H1 subsets rather than an exclusive role of these subsets during the development of autoimmune diseases and other tissue inflammation[34]. In this scenario, T_H17 cells might facilitate the migration of other T_H cells (such as T_H1 cells) into the target tissue, which could further propagate and modulate inflammation and tissue damage.

Taken together, these data suggest that T_H17 cells are potent inducers of autoimmunity through the promotion of tissue inflammation and the mobilization of the innate immune system. However, in some tissues, such as the gut and perhaps the liver[35], T_H17 cells, as potent early players of the adaptive immune system, might also have modulatory and protective roles.

At least two cytokines are needed to differentiate T_H17 cells

In contrast to T_H1 and T_H2 cell differentiation, which depend on their respective effector cytokines (IFN-γ and IL-4) for differentiation, T_H17 differentiation does not require IL-17. Instead, IL-6 and TGF-β—two cytokines with opposing effects—together induce the development of T_H17 cells[15,36,37]. IL-6 is a pro-inflammatory cytokine strongly induced in cells of the innate immune system on engagement

of specific pattern-recognition receptors such as Toll-like receptors and C-type lectin receptors. Thus, infection or local inflammation induces large amounts of IL-6. In the immune system, TGF-β is regarded as an anti-inflammatory cytokine because the loss of TGF-β is associated with a fatal lymphoproliferative disease[38]. In mice, TGF-β plus IL-6 have also been shown to be the differentiating factors for T_H17 cells *in vivo*. First, TGF-β transgenic animals immunized with the myelin oligodendrocyte glycoprotein $MOG_{35–55}$ in complete Freund's adjuvant, which induces high amounts of IL-6, develop exacerbated EAE owing to enhanced frequencies of T_H17 cells[37]. Second, mice with a defect in TGF-β responsiveness in T cells are protected from EAE owing to the lack of generation of T_H17 cells[39]. Third, when TGF-β is not secreted by T cells as a result of a conditionally disrupted *Tgfb* gene in CD4 cells, T_H17 cells cannot be generated and the mice are relatively protected from developing EAE[40]. Consistent with the idea that T_H17 cells require both TGF-β and IL-6, we showed that IL-6-deficient mice fail to develop a T_H17 response and are resistant to the development of EAE[10,37].

In humans, IL-17-producing T cells have been detected in the memory population of peripheral blood mononuclear cells (PBMCs); in one report they were characterized by the combined expression of CCR4 and CCR6 (ref. 41), and in another they were characterized by the expression of CCR2 and lack of CCR5 (ref. 42). A heterogeneous population of IL-17 and IFN-γ double-producers resided in the $CCR6^+CXCR3^+$ human memory-T-cell compartment[41]. It has been reported that in naive human T cells, the combination of TGF-β plus IL-6 or TGF-β plus IL-21 failed to induce the differentiation of T_H17 cells[43,44]. Recent findings suggested that $CD45RA^+$ human $CD4^+$ T cells can be more efficiently differentiated into T_H17 cells by a combination of IL-1β plus IL-6 (ref. 43) or IL-1β plus IL-23 (ref. 31). However, the presence of TGF-β in fetal calf serum or human serum used in these culture conditions cannot be totally excluded. Interestingly, another study indicated that the combination of TGF-β plus IL-6 is capable of inducing the expression of ROR-γt, a transcription factor important for T_H17 cells (see below), but not the expression of IL-17 in human T cells[45]. More recently, TGF-β in combination with IL-1β, IL-6 or IL-21 was shown to induce the differentiation of human T_H17 cells[99,100], thus highlighting the role of TGF-β in T_H17 differentiation. This also underscores the similarities in the differentiation of mouse and human T_H17 cells.

IL-21 and other cytokines as amplifiers of T_H17 cells

Although IFN-γ and IL-4 produced by T_H cells reinforce, T_H1 and T_H2 differentiation, respectively[46], IL-17 does not act on the differentiation and expansion of T_H17 cells. Three independent groups reported simultaneously that IL-21, a member of the IL-2 family of cytokines, is produced in overwhelming amounts by T_H17 cells and could, in combination with TGF-β, induce T_H17-differentiation[10–12]. Therefore, IL-21 produced by natural killer cells and natural killer T cells could induce the differentiation of T_H17 cells in the absence of IL-6 (ref. 10). When IL-6 is present, however, IL-21-receptor-deficient mice show a reduced but detectable T_H17 response *in vitro* and *in vivo*[10]. These findings point to a relevant function of IL-21, produced by newly generated T_H17 cells, in amplifying the precursor frequency of differentiating T_H17 cells (Fig. 2). In addition to IL-21, other cytokines such as tumour-necrosis factor alpha (TNF-α) and IL-1, which are not specifically produced by T_H17 cells, have been proposed to have an additional role in the amplification of T_H17 responses[36,47].

What is the role of IL-23?

The observation that IL-23p19-deficient animals, which did not develop EAE, lack IL-17-producing T cells[24] and the fact that IL-23 could expand a population of IL-17-producing pathogenic cells[3] pointed to an important role of IL-23 in the development of pathogenic T_H17 cells. It is now clear, however, that IL-23 does not act on naive T cells to induce their differentiation, but instead acts on already differentiated T_H17 cells. The maintenance of T_H17 cells *in vitro* for extended periods of time appears to require IL-23, which might also modulate effector functions of T_H17 cells both *in vitro* and *in vivo*. Recently it has been shown that T cells cultured in the presence of TGF-β plus IL-6 did not induce tissue inflammation unless they are further cultured in the presence of IL-23, which could decrease the secretion of IL-10 by these cells[48,49].

Although IL-23 was described eight years ago[50], little is known about the role of IL-23 for T_H17 cells *in vivo*. IL-23 signals through a receptor composed of the IL-12Rβ1 chain (which it shares with the IL-12 receptor) and a unique IL-23R subunit[51]. *IL23R* mRNA expression has mainly been detected in T cells, natural killer cells and natural killer T cells, but low levels of this receptor can also be found in monocytes, macrophages and dendritic cells[52]. Both IL-6 and IL-21 are strong inducers of IL-23R in T cells[12]. Furthermore, IL-23 can enhance the expression of its own receptor through an autocrine or paracrine feedback loop in mouse[3] (M. Oukka, unpublished observation) and human[45] T cells. The understanding of the regulation of IL-23R expression on naive T cells as they develop into T_H17 cells and on cells of the innate immune system will shed light on the role of this cytokine.

There is new evidence that IL-23 may have a profound impact on the innate immune system as well. Recent work has demonstrated that the development of gut inflammation in T-cell-deficient mice was dependent on IL-23, in that the loss of IL-23 but not IL-12 was associated with a decrease in gut inflammation mediated by anti-CD40-antibody-activated cells of the innate immune system[53]. IL-23 appears to induce IL-17, IL-1 and IL-6 from cells of the innate immune system[47,53]. Whether IL-23-mediated gut inflammation is entirely dependent on IL-17 produced by cells of the innate immune system has not been addressed. Consistent with this finding, a genome-wide scan revealed that particular SNPs in the coding sequence (rs11209026, c.1142G > A, p.Arg381Gln) of the *IL23R* gene conferred strong protection from Crohn's disease[54]. It has not been tested whether the different IL-23R variants affect the innate or adaptive immune systems or both.

We propose that the full differentiation of T_H17 cells requires three different steps: induction, amplification and stabilization/maintenance (Fig. 2). (1) The differentiation is initiated by the combined

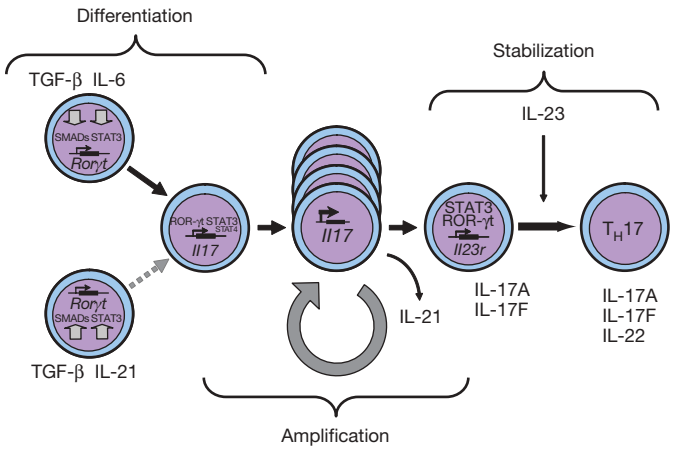


Figure 2 | Steps in the differentiation of T_H17 cells. Different factors control the initial differentiation of T_H17 cells from naive T cells, the amplification of T_H17 precursor cells, and finally the stabilization and effector phenotype of T_H17 cells. Whereas TGF-β together with IL-6 are the differentiation factors for T_H17 cells, IL-21, which is produced by T_H17 cells themselves, acts in a positive feedback loop to increase the frequency of T_H17 cells. STAT3 is the essential signalling molecule for the differentiation of T_H17 cells because the induction of IL-21 is absolutely dependent on STAT3, and STAT3 is also critical in the signal transduction cascades of IL-6, IL-21 and IL-23 receptors. IL-23 expands and stabilizes T_H17 cells to produce their effector cytokines IL-17, IL-17F and IL-22.

actions of IL-6 and TGF-β; (2) the amplification of the T_H17 response is driven through the production of IL-21 by T_H17 cells; (3) the stabilization/maintenance of the T_H17 phenotype is achieved by IL-23. Whereas the first two steps in the development of T_H17 cells seem to be distinct, it is possible that the stabilization and the amplification phases overlap or take place simultaneously.

T_H17-specific transcription factors

The differentiation of effector T_H cells is initiated by proximal signals from the T-cell receptor, co-stimulatory molecules and cytokine receptors. These integrated signals then lead to the induction of lineage-specific transcription factors. T_H17 cells have emerged as a truly independent subset because their differentiation was shown to be independent of the T_H1- or T_H2-promoting transcription factors T-bet, STAT1, STAT4 and STAT6 (refs 55–57). Consistent with the role of IL-6 in the differentiation of T_H17 cells, STAT3 appears to be critical for the differentiation of T_H17 cells because conditional deletion of STAT3 in T cells prevents the development of T_H17 cells[58]. Activation in the presence of TGF-β1 alone normally induces FOXP3 and generates induced T_{reg} cells (iT_{reg} cells). In T_{reg} cells, the interaction of TGF-β1 with its receptor induces the phosphorylation of SMAD2/3 proteins and the formation of a complex with SMAD4, which then translocates to the nucleus. Whether similar molecules are involved in the differentiation of T_H17 cells remains to be determined.

Analogously to T_H1 and T_H2 subsets, T_H17 development relies on the action of a lineage-specific transcription factor, identified as the orphan nuclear receptor ROR-γt. ROR-γt is selectively expressed in T_H17 cells differentiated in the presence of TGF-β plus IL-6, and transduction of naive T cells with a retroviral vector containing ROR-γt induces the development of T_H17 cells[59]. Conversely, loss of ROR-γt in T cells prevents the generation of myelin-specific T_H17 cells and subsequently the development of EAE in mice immunized with myelin antigens. Furthermore, the analysis of ROR-γt–GFP (green fluorescent protein) reporter mice revealed the existence of a population of IL-17^+ cells that are constitutively present in the intestinal lamina propria and whose development depends on ROR-γt expression[59]. These observations argue in favour of a critical role of ROR-γt in the differentiation of the T_H17 lineage (Fig. 2). However, the mechanisms by which ROR-γt drives T_H17 development have not yet been fully elucidated. A recent report indicates that, similarly to T-bet inducing IL-12Rβ2, ROR-γt would induce IL-23R. More precisely, IL-6-mediated activation of both STAT3 and ROR-γt would promote the production of IL-21 by T_H17 cells, and IL-21 would then induce the expression of IL-23R and establish responsiveness of T_H17 cells to IL-23. In this model, a sequential involvement of IL-6, IL-21 and IL-23 would lead to the differentiation of T_H17 cells[12]. Another report suggested that the induction of IL-21 required STAT3, but not ROR-γt[11]. In addition, it has not yet been formally addressed whether ROR-γt directly transactivates the *IL17A*, *IL17F*, *IL22* and *IL21* genes in T_H17 cells. A more recent report suggested that T_H17 differentiation might be mediated by the combined effect of ROR-γt and ROR-α, both of which are expressed at high levels in differentiated T_H17 cells. Loss of only one of these transcription factors resulted in partial loss of T_H17 cytokine expression, and loss of both ROR-γt and ROR-α abrogated T_H17 differentiation[60].

Reciprocal relationship between iT_{reg} and T_H17 cells

TGF-β induces the T_{reg}-specific transcription factor FOXP3 and is required for the maintenance of iT_{reg} cells in the peripheral immune compartment. However, addition of IL-6 to TGF-β inhibits the generation of T_{reg} cells and induces T_H17 cells. On the basis of these data, we first proposed[37] that there is a reciprocal relationship between T_{reg} cells and T_H17 cells, and that IL-6 has a pivotal role in dictating the balance between these two cell populations[10,37].

This reciprocal relationship between T_{reg} and T_H17 cells is further supported by recent data from other laboratories[61,62]: IL-2, which is a growth factor for T_{reg} cells, has also been shown to inhibit the generation of T_H17 cells[61]. Consistent with these data, mice that lack IL-2 or in which IL-2 signalling is compromised (*Stat5*$^{-/-}$), harbour reduced numbers of T_{reg} cells and an increased proportion of T_H17 cells in the peripheral repertoire[61]. Moreover, these mice develop multi-organ inflammatory diseases, which can be prevented by the passive transfer of T_{reg} cells[63].

Additional evidence for a reciprocal developmental relationship between FOXP3$^+$ T_{reg} cells and T_H17 cells came from the discovery that retinoic acid, a vitamin A metabolite, could drive the generation of T_{reg} cells[64] by enhancing TGF-β signalling and enhancing FOXP3 promoter activity while abrogating the differentiation of T_H17 cells, but not of T_H1 cells, through the inhibition of IL-6 signalling[62]. These findings indicate that retinoic acid can regulate the balance between pro-inflammatory T_H17 cells and anti-inflammatory T_{reg} cells (Fig. 3).

Finally, it has been found that ROR-γt and RORα, the transcription factors for T_H17 cells, and FOXP3, the transcription factor for T_{reg} cells, can physically bind to each other and antagonize each other's functions[65,66]. In line with this concept, conditional deletion of FOXP3 protein in 'T_{reg} cells' *in vivo* resulted in an increase in ROR-γt, IL-17 and IL-21 expression[67,68], further corroborating the reciprocal relationship between T_H17 cells and T_{reg} cells.

More than one way to inhibit T_H17 cells

So far, several cytokines and pathways have been shown to inhibit the development and expansion of T_H17 cells. T_H1- and T_H2-specific cytokines can antagonize each other. Correspondingly, IL-4, IL-25 (IL-17E) and IFN-γ also inhibit the expansion of T_H17 cells[55,56,69].

Similarly, IL-27, a member of the IL-12 heterodimeric family of cytokines produced by cells of the innate immune system, can suppress the development of T_H17 responses. Consistent with this observation, a lack of IL-27 signalling resulted in an increased T_H17

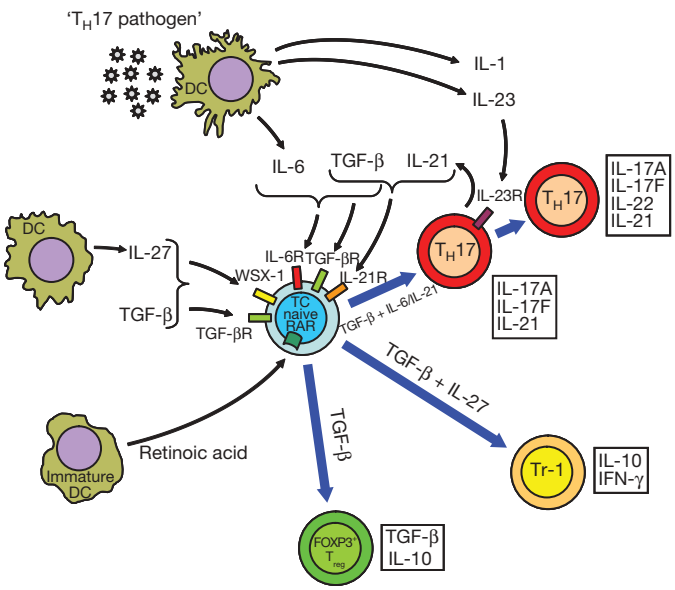


Figure 3 | The developmental pathways of T_H17 cells and FOXP3$^+$ T_{reg} cells require TGF-β signalling and are reciprocally regulated. TGF-β is ubiquitous although its most relevant source in regulating immune reactions is still unclear. Other factors such as retinoic acid or cytokines such as IL-6, IL-1, IL-23 or IL-27 are provided by cells of the innate immune system (immature or activated dendritic cells (DCs), respectively) and dictate whether a naive T cell (TC) develops into a FOXP3$^+$ T_{reg} cell, a T_H17 cell or an IL-10-secreting T cell of the Tr-1 phenotype. IL-6R, IL-6 receptor; IL-21R, IL-21 receptor; IL-23R, IL-23 receptor; RAR, retinoic acid receptor; TGF-βR, TGF-β receptor; WSX-1, IL-27 receptor).

response and enhanced inflammation of the central nervous system in two different disease models[70,71]. Two recent studies showed that IL-27 together with TGF-β induces the differentiation of IL-10-producing T cells with Tr-1-like properties and that IL-27R-deficient mice (*Wsx1*$^{-/-}$) have a defect in generating IL-10-producing Tr-1 cells[48,72]. Thus, IL-27 might also be necessary to control exaggerated immunopathology indirectly by inducing Tr-1 cells (Fig. 3).

Role of TGF-β in inducing novel T_H subsets

Since the discovery of TGF-β and IL-6 as the differentiation factors for T_H17 cells, we have proposed that the dual cytokine interaction (TGF-β plus a cytokine X) might be operational in the induction of other novel T-cell subsets as well. Tr-1 differentiation induced by a combination of TGF-β plus IL-27 supports this hypothesis[48,72]. When TGF-β is involved in the differentiation of novel T_H subsets, T_H commitment might be accomplished either by TGF-β acting together with another cytokine, where the two cytokines will inhibit each other's functions and result in the generation of a totally new gene programme (for example, T_H17 differentiation induced by TGF-β plus IL-6), or by quantitatively scaling back each other's functions, thereby resulting in the production of only dominant cytokines in the responding T cells (for example, Tr-1 cells induced by TGF-β plus IL-27; Fig. 4). These observations suggest that T cells would sense multiple cytokine inputs simultaneously from the environment to initiate the differentiation of new T-cell subsets with distinct cytokine phenotypes and effector functions.

Concluding remarks

It is now established that T_H17 cells constitute an independent T-helper-cell subset with major functions in the induction of tissue inflammation and host protection against extracellular pathogens. Since their initial description, substantial progress has been made in the understanding of T_H17 differentiation and effector functions. On the basis of recent reports, we propose a three-step model for the differentiation of T_H17 cells: induction, amplification and maintenance/stabilization, where TGF-β plus IL-6 induce the differentiation of T_H17 cells, IL-21 amplifies the frequency of T_H17 cells and IL-23 stabilizes the phenotype of previously differentiated T_H17 cells. Loss of any one of the members in this pathway (IL-6, IL-21 or IL-23) limits the T_H17 response, and only the combination of these factors leads to a robust and stable T_H17 response. The understanding of how different cytokine signalling pathways are integrated to induce the differentiation of novel T_H subsets, including T_H17 cells, will represent a major step forward in our understanding of T-cell-subset differentiation. Because multiple lines of evidence suggest that there is a reciprocal relationship between T_{reg} cells and T_H17 cells, manipulation of this differentiation pathway might result in the generation of pro-inflammatory T_H17 cells and induce tissue inflammation or induce protective T_{reg} cells and therefore inhibit autoimmunity and induce tolerance. Targeting nodal points in this pathway will allow one to shift the balance between pro-inflammatory T_H17 cells and inhibitory T_{reg} cells and thus provide exciting new targets for the treatment of multiple inflammatory and autoimmune diseases.

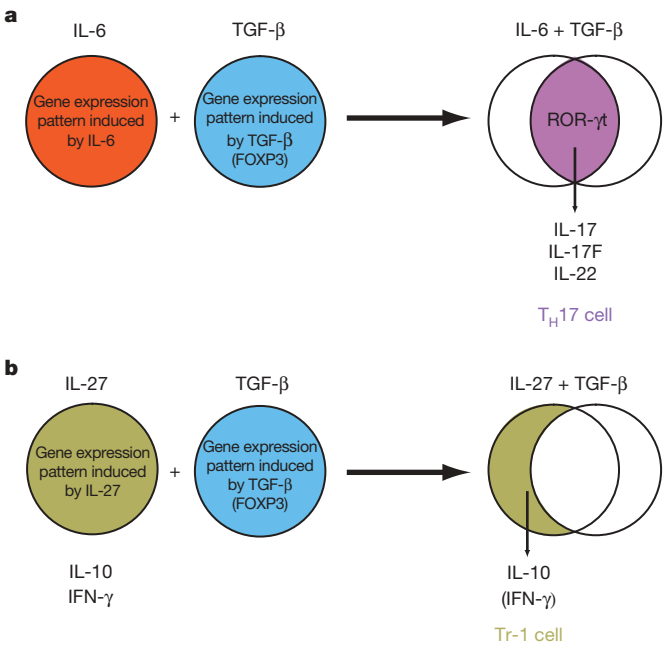


Figure 4 | Effects of TGF-β in shaping the transcriptional programme of developing T helper cell subsets. a, IL-6 and TGF-β independently induce specific gene expression programmes in T cells. However, when both cytokines act in concert, an essentially novel and distinct gene expression programme is induced resulting in a qualitatively different outcome such as the T_H17 transcriptional programme *(IL-17*, *IL-17F* and *IL-22*). **b**, In contrast, when TGF-β acts in combination with cytokines such as IL-27, IFN-γ expression is scaled down and IL-10 expression is increased resulting in a Tr-1-like phenotype.

1. Mosmann, T. R. & Coffman, R. L. T_H1 and T_H2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* **7**, 145–173 (1989).
2. Fort, M. M. *et al.* IL-25 induces IL-4, IL-5, and IL-13 and Th2-associated pathologies *in vivo*. *Immunity* **15**, 985–995 (2001).
3. Langrish, C. L. *et al.* IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. *J. Exp. Med.* **201**, 233–240 (2005).
4. Kolls, J. K. & Linden, A. Interleukin-17 family members and inflammation. *Immunity* **21**, 467–476 (2004).
5. Gunimaladevi, I., Savan, R. & Sakai, M. Identification, cloning and characterization of interleukin-17 and its family from zebrafish. *Fish Shellfish Immunol.* **21**, 393–403 (2006).
6. Liang, S. C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203**, 2271–2279 (2006).
7. Chang, S. H. & Dong, C. A novel heterodimeric cytokine consisting of IL-17 and IL-17F regulates inflammatory responses. *Cell Res.* **17**, 435–440 (2007).
8. Liang, S. C. *et al.* An IL-17F/A heterodimer protein is produced by mouse Th17 cells and induces airway neutrophil recruitment. *J. Immunol.* **179**, 7791–7799 (2007).
9. Zheng, Y. *et al.* Interleukin-22, a T_H17 cytokine, mediates IL-23-induced dermal inflammation and acanthosis. *Nature* **445**, 648–651 (2007).
10. Korn, T. *et al.* IL-21 initiates an alternative pathway to induce proinflammatory T_H17 cells. *Nature* **448**, 484–487 (2007).
11. Nurieva, R. *et al.* Essential autocrine regulation by IL-21 in the generation of inflammatory T cells. *Nature* **448**, 480–483 (2007).
12. Zhou, L. *et al.* IL-6 programs T_H-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nature Immunol.* **8**, 967–974 (2007).
13. Chtanova, T. *et al.* T follicular helper cells express a distinctive transcriptional profile, reflecting their role as non-Th1/Th2 effector cells that provide help for B cells. *J. Immunol.* **173**, 68–78 (2004).
14. Ye, P. *et al.* Requirement of interleukin 17 receptor signaling for lung CXC chemokine and granulocyte colony-stimulating factor expression, neutrophil recruitment, and host defense. *J. Exp. Med.* **194**, 519–527 (2001).
15. Mangan, P. R. *et al.* Transforming growth factor-β induces development of the T_H17 lineage. *Nature* **441**, 231–234 (2006).
16. Chung, D. R. *et al.* $CD4^+$ T cells mediate abscess formation in intra-abdominal sepsis by an IL-17-dependent mechanism. *J. Immunol.* **170**, 1958–1963 (2003).
17. Infante-Duarte, C., Horton, H. F., Byrne, M. C. & Kamradt, T. Microbial lipopeptides induce the production of IL-17 in Th cells. *J. Immunol.* **165**, 6107–6115 (2000).
18. Khader, S. A. *et al.* IL-23 and IL-17 in the establishment of protective pulmonary $CD4^+$ T cell responses after vaccination and during *Mycobacterium tuberculosis* challenge. *Nature Immunol.* **8**, 369–377 (2007).
19. Huang, W., Na, L., Fidel, P. L. & Schwarzenberger, P. Requirement of interleukin-17A for systemic anti-*Candida albicans* host defense in mice. *J. Infect. Dis.* **190**, 624–631 (2004).
20. van Beelen, A. J. *et al.* Stimulation of the intracellular bacterial sensor NOD2 programs dendritic cells to promote interleukin-17 production in human memory T cells. *Immunity* **27**, 660–669 (2007).
21. Charlton, B. & Lafferty, K. J. The Th1/Th2 balance in autoimmunity. *Curr. Opin. Immunol.* **7**, 793–798 (1995).
22. Ferber, I. A. *et al.* Mice with a disrupted IFN-γ gene are susceptible to the induction of experimental autoimmune encephalomyelitis (EAE). *J. Immunol.* **156**, 5–7 (1996).
23. Becher, B., Durell, B. G. & Noelle, R. J. Experimental autoimmune encephalitis and inflammation in the absence of interleukin-12. *J. Clin. Invest.* **110**, 493–497 (2002).

24. Cua, D. J. *et al.* Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain. *Nature* **421,** 744–748 (2003).
25. Sato, K. *et al.* Th17 functions as an osteoclastogenic helper T cell subset that links T cell activation and bone destruction. *J. Exp. Med.* **203,** 2673–2682 (2006).
26. Nakae, S., Nambu, A., Sudo, K. & Iwakura, Y. Suppression of immune induction of collagen-induced arthritis in IL-17-deficient mice. *J. Immunol.* **171,** 6173–6177 (2003).
27. Komiyama, Y. *et al.* IL-17 plays an important role in the development of experimental autoimmune encephalomyelitis. *J. Immunol.* **177,** 566–573 (2006).
28. Chabaud, M. *et al.* Human interleukin-17: a T cell-derived proinflammatory cytokine produced by the rheumatoid synovium. *Arthritis Rheum.* **42,** 963–970 (1999).
29. Lock, C. *et al.* Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nature Med.* **8,** 500–508 (2002).
30. Fujino, S. *et al.* Increased expression of interleukin 17 in inflammatory bowel disease. *Gut* **52,** 65–70 (2003).
31. Wilson, N. J. *et al.* Development, cytokine profile and function of human interleukin 17-producing helper T cells. *Nature Immunol.* **8,** 950–957 (2007).
32. Kebir, H. *et al.* Human T_H17 lymphocytes promote blood–brain barrier disruption and central nervous system inflammation. *Nature Med.* **13,** 1173–1175 (2007).
33. Ben-Nun, A., Wekerle, H. & Cohen, I. R. The rapid isolation of clonable antigen-specific T lymphocyte lines capable of mediating autoimmune encephalomyelitis. *Eur. J. Immunol.* **11,** 195–199 (1981).
34. Lohr, J., Knoechel, B., Wang, J. J., Villarino, A. V. & Abbas, A. K. Role of IL-17 and regulatory T lymphocytes in a systemic autoimmune disease. *J. Exp. Med.* **203,** 2785–2791 (2006).
35. Zenewicz, L. A. *et al.* Interleukin-22 but not interleukin-17 provides protection to hepatocytes during acute liver inflammation. *Immunity* **27,** 647–659 (2007).
36. Veldhoen, M., Hocking, R. J., Atkins, C. J., Locksley, R. M. & Stockinger, B. TGFβ in the context of an inflammatory cytokine milieu supports *de novo* differentiation of IL-17-producing T cells. *Immunity* **24,** 179–189 (2006).
37. Bettelli, E. *et al.* Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells. *Nature* **441,** 235–238 (2006).
38. Kulkarni, A. B. *et al.* Transforming growth factor beta 1 null mutation in mice causes excessive inflammatory response and early death. *Proc. Natl Acad. Sci. USA* **90,** 770–774 (1993).
39. Veldhoen, M., Hocking, R. J., Flavell, R. A. & Stockinger, B. Signals mediated by transforming growth factor-beta initiate autoimmune encephalomyelitis, but chronic inflammation is needed to sustain disease. *Nature Immunol.* **7,** 1151–1156 (2006).
40. Li, M. O., Wan, Y. Y. & Flavell, R. A. T. Cell-produced transforming growth factor-β1 controls T cell tolerance and regulates Th1- and Th17-cell differentiation. *Immunity* **26,** 579–591 (2007).
41. Acosta-Rodriguez, E. V. *et al.* Surface phenotype and antigenic specificity of human interleukin 17-producing T helper memory cells. *Nature Immunol.* **8,** 639–646 (2007).
42. Sato, W., Aranami, T. & Yamamura, T. Cutting edge: human Th17 cells are identified as bearing $CCR2^+CCR5^-$ phenotype. *J. Immunol.* **178,** 7525–7529 (2007).
43. Acosta-Rodriguez, E. V., Napolitani, G., Lanzavecchia, A. & Sallusto, F. Interleukins 1β and 6 but not transforming growth factor-β are essential for the differentiation of interleukin 17-producing human T helper cells. *Nature Immunol.* **8,** 942–949 (2007).
44. Evans, H. G., Suddason, T., Jackson, I., Taams, L. S. & Lord, G. M. Optimal induction of T helper 17 cells in humans requires T cell receptor ligation in the context of Toll-like receptor-activated monocytes. *Proc. Natl Acad. Sci. USA* **104,** 17034–17039 (2007).
45. Chen, Z., Tato, C. M., Muul, L., Laurence, A. & O'Shea, J. J. Distinct regulation of interleukin-17 in human T helper lymphocytes. *Arthritis Rheum.* **56,** 2936–2946 (2007).
46. Murphy, K. M. & Reiner, S. L. The lineage decisions of helper T cells. *Nature Rev. Immunol.* **2,** 933–944 (2002).
47. Sutton, C., Brereton, C., Keogh, B., Mills, K. H. & Lavelle, E. C. A crucial role for interleukin (IL)-1 in the induction of IL-17-producing T cells that mediate autoimmune encephalomyelitis. *J. Exp. Med.* **203,** 1685–1691 (2006).
48. Stumhofer, J. S. *et al.* Interleukins 27 and 6 induce STAT3-mediated T cell production of interleukin 10. *Nature Immunol.* **8,** 1363–1371 (2007).
49. McGeachy, M. J. *et al.* TGF-β and IL-6 drive the production of IL-17 and IL-10 by T cells and restrain T_H-17 cell-mediated pathology. *Nature Immunol.* **8,** 1390–1397 (2007).
50. Oppmann, B. *et al.* Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity* **13,** 715–725 (2000).
51. Kastelein, R. A., Hunter, C. A. & Cua, D. J. Discovery and biology of IL-23 and IL-27: related but functionally distinct regulators of inflammation. *Annu. Rev. Immunol.* **25,** 221–242 (2007).
52. Parham, C. *et al.* A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rβ1 and a novel cytokine receptor subunit, IL-23R. *J. Immunol.* **168,** 5699–5708 (2002).
53. Uhlig, H. H. *et al.* Differential activity of IL-12 and IL-23 in mucosal and systemic innate immune pathology. *Immunity* **25,** 309–318 (2006).
54. Duerr, R. H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314,** 1461–1463 (2006).
55. Park, H. *et al.* A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunol.* **6,** 1133–1141 (2005).
56. Harrington, L. E. *et al.* Interleukin 17-producing $CD4^+$ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nature Immunol.* **6,** 1123–1132 (2005).
57. Chen, Z. *et al.* Selective regulatory function of Socs3 in the formation of IL-17-secreting T cells. *Proc. Natl Acad. Sci. USA* **103,** 8137–8142 (2006).
58. Yang, X. O. *et al.* STAT3 regulates cytokine-mediated generation of inflammatory helper T cells. *J. Biol. Chem.* **282,** 9358–9363 (2007).
59. Ivanov, I. I. *et al.* The orphan nuclear receptor RORγt directs the differentiation program of proinflammatory $IL\text{-}17^+$ T helper cells. *Cell* **126,** 1121–1133 (2006).
60. Yang, X. O. *et al.* T helper 17 lineage differentiation is programmed by orphan nuclear receptors RORα and RORγ. *Immunity* **28,** 29–39 (2008).
61. Laurence, A. *et al.* Interleukin-2 signaling via STAT5 constrains T helper 17 cell generation. *Immunity* **26,** 371–381 (2007).
62. Mucida, D. *et al.* Reciprocal T_H17 and regulatory T cell differentiation mediated by retinoic acid. *Science* **317,** 256–260 (2007).
63. Antov, A., Yang, L., Vig, M., Baltimore, D. & Van Parijs, L. Essential role for STAT5 signaling in $CD25^+CD4^+$ regulatory T cell homeostasis and the maintenance of self-tolerance. *J. Immunol.* **171,** 3435–3441 (2003).
64. Coombes, J. L. *et al.* A functionally specialized population of mucosal $CD103^+$ DCs induces $Foxp3^+$ regulatory T cells via a TGF-β and retinoic acid-dependent mechanism. *J. Exp. Med.* **204,** 1757–1764 (2007).
65. Zhou, L. *et al.* TGF-β-induced Foxp3 inhibits T_H17 cell differentiation by antagonizing RORγt function. *Nature* **453,** 236–240 (2008).
66. Du, J., Huang, C., Zhou, B. & Ziegler, S. F. Isoform-specific inhibition of RORα-mediated transcriptional activation by human FOXP3. *J. Immunol.* **180,** 4785–4792 (2008).
67. Gavin, M. A. *et al.* Foxp3-dependent programme of regulatory T-cell differentiation. *Nature* **445,** 771–775 (2007).
68. Williams, L. M. & Rudensky, A. Y. Maintenance of the Foxp3-dependent developmental program in mature regulatory T cells requires continued expression of Foxp3. *Nature Immunol.* **8,** 277–284 (2007).
69. Kleinschek, M. A. *et al.* IL-25 regulates Th17 function in autoimmune inflammation. *J. Exp. Med.* **204,** 161–170 (2007).
70. Batten, M. *et al.* Interleukin 27 limits autoimmune encephalomyelitis by suppressing the development of interleukin 17-producing T cells. *Nature Immunol.* **7,** 929–936 (2006).
71. Stumhofer, J. S. *et al.* Interleukin 27 negatively regulates the development of interleukin 17-producing T helper cells during chronic inflammation of the central nervous system. *Nature Immunol.* **7,** 937–945 (2006).
72. Awasthi, A. *et al.* A dominant function for interleukin 27 in generating interleukin 10-producing anti-inflammatory T cells. *Nature Immunol.* **8,** 1380–1389 (2007).
73. Liu, S. J. *et al.* Induction of a distinct CD8 Tnc17 subset by transforming growth factor-β and interleukin-6. *J. Leukoc. Biol.* **82,** 354–360 (2007).
74. Lockhart, E., Green, A. M. & Flynn, J. L. IL-17 production is dominated by γδ T cells rather than CD4 T cells during *Mycobacterium tuberculosis* infection. *J. Immunol.* **177,** 4662–4669 (2006).
75. Ferretti, S., Bonneau, O., Dubois, G. R., Jones, C. E. & Trifilieff, A. IL-17, produced by lymphocytes and neutrophils, is necessary for lipopolysaccharide-induced airway neutrophilia: IL-15 as a possible trigger. *J. Immunol.* **170,** 2106–2112 (2003).
76. Molet, S. *et al.* IL-17 is increased in asthmatic airways and induces human bronchial fibroblasts to produce cytokines. *J. Allergy Clin. Immunol.* **108,** 430–438 (2001).
77. Zhou, Q., Desta, T., Fenton, M., Graves, D. T. & Amar, S. Cytokine profiling of macrophages exposed to *Porphyromonas gingivalis*, its lipopolysaccharide, or its FimA protein. *Infect. Immun.* **73,** 935–943 (2005).
78. Yao, Z. *et al.* Herpesvirus Saimiri encodes a new cytokine, IL-17, which binds to a novel cytokine receptor. *Immunity* **3,** 811–821 (1995).
79. Kuestner, R. E. *et al.* Identification of the IL-17 receptor related molecule IL-17RC as the receptor for IL-17F. *J. Immunol.* **179,** 5462–5473 (2007).
80. Toy, D. *et al.* Cutting edge: interleukin 17 signals through a heteromeric receptor complex. *J. Immunol.* **177,** 36–39 (2006).
81. Fossiez, F. *et al.* T cell interleukin-17 induces stromal cells to produce proinflammatory and hematopoietic cytokines. *J. Exp. Med.* **183,** 2593–2603 (1996).
82. Martel-Pelletier, J., Mineau, F., Jovanovic, D., Di Battista, J. A. & Pelletier, J. P. Mitogen-activated protein kinase and nuclear factor κB together regulate interleukin-17-induced nitric oxide production in human osteoarthritic chondrocytes: possible role of transactivating factor mitogen-activated protein kinase-activated proten kinase (MAPKAPK). *Arthritis Rheum.* **42,** 2399–2409 (1999).
83. Stark, M. A. *et al.* Phagocytosis of apoptotic neutrophils regulates granulopoiesis via IL-23 and IL-17. *Immunity* **22,** 285–294 (2005).
84. Starnes, T. *et al.* Cutting edge: IL-17F, a novel cytokine selectively expressed in activated T cells and monocytes, regulates angiogenesis and endothelial cell cytokine production. *J. Immunol.* **167,** 4137–4140 (2001).
85. Hymowitz, S. G. *et al.* IL-17s adopt a cystine knot fold: structure and activity of a novel cytokine, IL-17F, and implications for receptor binding. *EMBO J.* **20,** 5332–5341 (2001).

86. Hurst, S. D. *et al.* New IL-17 family members promote Th1 or Th2 responses in the lung: *in vivo* function of the novel cytokine IL-25. *J. Immunol.* **169,** 443–453 (2002).
87. Wolk, K. & Sabat, R. Interleukin-22: a novel T- and NK-cell derived cytokine that regulates the biology of tissue cells. *Cytokine Growth Factor Rev.* **17,** 367–380 (2006).
88. Kotenko, S. V. *et al.* Identification of the functional interleukin-22 (IL-22) receptor complex: the IL-10R2 chain (IL-10Rβ) is a common chain of both the IL-10 and IL-22 (IL-10-related T cell-derived inducible factor, IL-TIF) receptor complexes. *J. Biol. Chem.* **276,** 2725–2732 (2001).
89. Moore, K. W., de Waal Malefyt, R., Coffman, R. L. & O'Garra, A. Interleukin-10 and the interleukin-10 receptor. *Annu. Rev. Immunol.* **19,** 683–765 (2001).
90. Wolk, K. *et al.* IL-22 increases the innate immunity of tissues. *Immunity* **21,** 241–254 (2004).
91. Dumoutier, L., Van Roost, E., Colau, D. & Renauld, J. C. Human interleukin-10-related T cell-derived inducible factor: molecular cloning and functional characterization as an hepatocyte-stimulating factor. *Proc. Natl Acad. Sci. USA* **97,** 10144–10149 (2000).
92. Parrish-Novak, J. *et al.* Interleukin 21 and its receptor are involved in NK cell expansion and regulation of lymphocyte function. *Nature* **408,** 57–63 (2000).
93. Leonard, W. J. & Spolski, R. Interleukin-21: a modulator of lymphoid proliferation, apoptosis and differentiation. *Nature Rev. Immunol.* **5,** 688–698 (2005).
94. Takeshita, T. *et al.* Cloning of the gamma chain of the human IL-2 receptor. *Science* **257,** 379–382 (1992).
95. Zeng, R. *et al.* Synergy of IL-21 and IL-15 in regulating $CD8^+$ T cell expansion and function. *J. Exp. Med.* **201,** 139–148 (2005).
96. Spolski, R. & Leonard, W. J. Interleukin-21: basic biology and implications for cancer and autoimmunity. *Annu. Rev. Immunol.* **26,** 57–79 (2008).
97. Coquet, J. M. *et al.* IL-21 is produced by NKT cells and modulates NKT cell activation and cytokine production. *J. Immunol.* **178,** 2827–2834 (2007).
98. Pelletier, M., Bouchard, A. & Girard, D. *In vivo* and *in vitro* roles of IL-21 in inflammation. *J. Immunol.* **173,** 7521–7530 (2004).
99. Manel, N., Unutmaz, D. & Littman, D. R. The differentiation of human T(H)-17 cells requires transforming growth factor-beta and induction of the nuclear receptor RORgamma. *Nat. Immunol.* advance online publication doi:10.1038/ni.1610 (4 May 2008).
100. Yang, L. *et al.* IL-21 and TGF-β are required for differentiation of human T_H17 cells. *Nature* advance online publication doi:10.1038/nature07021 (11 May 2008).

Acknowledgements This work was supported by grants from the National Multiple Sclerosis Society, the National Institutes of Health, the Juvenile Diabetes Research Foundation Center for Immunological Tolerance at Harvard, and the Deutsche Forschungsgemeinschaft. V.K.K. is the recipient of the Javits Neuroscience Investigator Award from the National Institutes of Health.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to V.K.K. (vkuchroo@rics.bwh.harvard.edu).

ARTICLES

Micro-engineered local field control for high-sensitivity multispectral MRI

Gary Zabow[1,2], Stephen Dodd[1], John Moreland[2] & Alan Koretsky[1]

In recent years, biotechnology and biomedical research have benefited from the introduction of a variety of specialized nanoparticles whose well-defined, optically distinguishable signatures enable simultaneous tracking of numerous biological indicators. Unfortunately, equivalent multiplexing capabilities are largely absent in the field of magnetic resonance imaging (MRI). Comparable magnetic-resonance labels have generally been limited to relatively simple chemically synthesized superparamagnetic microparticles that are, to a large extent, indistinguishable from one another. Here we show how it is instead possible to use a top-down microfabrication approach to effectively encode distinguishable spectral signatures into the geometry of magnetic microstructures. Although based on different physical principles from those of optically probed nanoparticles, these geometrically defined magnetic microstructures permit a multiplexing functionality in the magnetic resonance radio-frequency spectrum that is in many ways analogous to that permitted by quantum dots in the optical spectrum. Additionally, *in situ* modification of particle geometries may facilitate radio-frequency probing of various local physiological variables.

Magnetic resonance imaging[1,2] has become a widely used medical diagnostic and research tool[3]. A key to this success has been the development of numerous chemically synthesized image-enhancing agents[4–8]. Nevertheless, MRI still lacks the sensitivity and the multiplexing capabilities of optical imaging, which can use coloured fluorophores[9], multi-spectral semiconductor quantum dots[10–12], metallic nanoparticles[13,14], and even microfabricated barcodes[15] for multifunctional encoding and biomolecular or cellular labelling, sensing and tracking. Because optically based labels can probe only so far beneath most surfaces, however, being able to distinguish with MRI between different types of cells, at the single-cell level, would be useful for cellular biology and early disease detection and diagnosis. Currently, however, MRI cell tracking is based on the magnetically dephased signal from the water surrounding cells labelled with many superparamagnetic iron oxide (SPIO) nanoparticles[6,16,17] or dendrimers[18], or individual micrometre-sized particles of iron oxide[19–21] (MPIOs). The continuous spatial decay of the external fields surrounding these, or any other, magnetizable particles imposes a continuous range of Larmor frequencies that broadens the water hydrogen proton line, obscuring distinction between different types of magnetic particles that might specifically label different types of cells. The utility of magnetic particles would be greatly enhanced if they could instead frequency-shift the water by discrete, controllable amounts, transforming an effectively monochrome contrast agent into a 'coloured' spectral set of distinguishable tags.

Here we consider the advantages of top-down microfabrication as an alternative to traditional bottom-up chemical synthesis for designing MRI contrast agents with more directly engineered properties and, accordingly, increased functionality. In particular, we demonstrate a new contrast agent imaging modality based on geometrical rather than chemical structure, showing how engineered magnetic microstructures can form effectively subcellular-sized radio-frequency identification (RFID) tags for multi-spectral MRI. Designed to exploit water diffusion, these microstructures locally increase MRI sensitivity by several orders of magnitude, yielding low concentration requirements and potentially enabling individually detectable, spectrally distinct micro-tags. With frequencies determined by structural shape and composition instead of by chemical[7] or nuclear[8] shift, spectral signatures can be tailored over very broad frequency shift ranges spanning many tens of thousands of parts per million. Beyond their radio-frequency analogy to continuously tunable optical quantum dots, such microstructures may also enable a variety of localized physiological probes, enhancing both MRI capabilities and basic biological research.

The potential of spectral shifting is indicated by recent interest in PARACEST[7] molecular complexes, whose chemical shifts can generate off-resonance MRI contrast through proton exchange. Unfortunately, relying on a restricted set of macromolecular structures, they have relatively limited shift ranges which may restrict their effectiveness and often necessitate high MRI fields. Here it is shown that instead of being constrained by any inherent chemical environment, it is possible to customize spectral shifts by microfabricating suitably shaped magnetizable elements. This increased design control allows shift ranges and sensitivities that far exceed those of existing molecular analogues, and enables a new class of MRI agent: single-particle spectral tags, which combine the advantages of single-particle tracking and distinct spectral shifting, while retaining compatibility with standard MRI hardware.

Magnetic structure design and operation

Spectral shifting by magnetic structures is made possible by noting that even though all magnetic objects have continuously decaying external fields, this does not preclude discretely frequency shifting MRI-detectable nuclei contained internally, either within a magnetizable shell or between neighbouring magnetizable elements. A distinct, resolvable frequency-shifted peak requires a spatially extended volume over which the additional field generated by the magnetizable structure is homogeneous and preferably offset in magnitude from that of the structure's surrounding external decaying fields. More precisely, because typical background MRI field magnitudes generally

[1]Laboratory of Functional and Molecular Imaging, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Electromagnetics Division, National Institute of Standards and Technology, Boulder, Colorado 80305, USA.

substantially exceed those of the magnetizable structures (at least in the regions of interest), quadrature addition implies that only that component of the structure's field parallel to the background field need strictly satisfy this homogeneity condition. Among several possible configurations that may be useful, we demonstrate here a spaced, magnetizable double-disk geometry that is illustrated schematically in Fig. 1, together with typical resulting magnetic field profiles. This particular geometry is attractive because, in addition to generating a highly homogeneous field over a large volume fraction, the particularly open nature of the design helps maximize water self-diffusion through the structure, enabling use of MRI techniques that can increase the signal-to-noise ratio over that of any closed structure.

The double-disk geometry is also inherently scalable and suited to parallel wafer-level microfabrication. Figure 2a–d shows scanning electron microscope (SEM) images of sample microfabricated structures. Full fabrication details are lengthy (G.Z., manuscript in preparation); briefly, particle complexes are surface micromachined through a combination of metal evaporation and electroplating depositions followed by lithographically defined ion-milling and selective wet etching. The disks are separated by non-magnetic spacers: either an internal metal post that remains after a timed etch, or external biocompatible[22,23] photo-epoxy posts. A final gold sputter-coating further enhances biocompatibility and access to thiol-based chemistry for specific surface functionalization if desired.

Although the structure's exact resonance frequency shift, $\Delta\omega$, depends on the fields generated throughout the volume between the disks, $\Delta\omega$ can be roughly approximated analytically from the field at the centre of the structure. For gyromagnetic ratio γ, and magnetically saturated disks of thickness h, radius R, centre-to-centre separation $2S$, and saturation magnetic polarization J_S, elementary magnetostatics gives $\Delta\omega = (\gamma J_S/2)[(S-h/2)/((S-h/2)^2+R^2)^{1/2} - (S+h/2)/((S+h/2)^2+R^2)^{1/2}]$. For thin disks with $h \ll 2S \approx R$, this reduces to:

$$\Delta\omega \approx -\gamma J_S \left(\frac{hR^2}{2(R^2+S^2)^{3/2}} \right)$$

Spectral signatures can be tailored by modifying any of J_S, h, R, or S. All particles shown in this Article were made from nickel ($J_S \approx 0.5$–0.6 T), but could equally well be formed from other magnetic alloys. J_S can therefore be chosen anywhere from zero up to 2 T (soft iron), enabling large water shift ranges from 0 to of order of -10 MHz. Unlike frequency shifting based on chemical molecules, the frequency dependence on a dimensionless geometrical aspect ratio implies shifting of any nuclear species and by any overall particle size. For example, we demonstrate here frequency shifting of both hydrogen (Figs 3 and 4b–e) and deuterium (Fig. 4a) nuclei, and with particle size scales spanning three orders of magnitude, from millimetre to micrometre.

This frequency-shifting ability implicitly assumes alignment of the disk planes with the applied magnetizing MRI field, B_0. Such alignment is ensured by the structure's built-in magnetic shape anisotropy. Figure 2e demonstrates this, showing particles readily self-aligning even in small fields. Although aligning torques generally increase with increasing field, once typical MRI fields are reached, the structures' magnetic materials are already fully saturated and their Zeeman magnetostatic energies are therefore independent of particle orientation. In this regime, aligning torque magnitudes decouple from B_0 and are instead determined by the angular dependence of the magnetostatic demagnetization energy[24,25] that is proportional to J_S^2/μ_0, for free-space magnetic permeability, $\mu_0 = 4\pi \times 10^{-7}$ H m^{-1}. Specifically, assuming $h \ll R$, for any misalignment angles θ between B_0 and the disk planes, resulting magnetic torques on the disks produce self-aligning pressures of order $(h/(R^2+S^2)^{1/2})(J_S^2/\mu_0)\sin(2\theta)$. This equates to pressures of order 10^{-8} to 10^{-6} N μm^{-2}. By comparison, even within cellular cytoplasm, yield stresses are only in the range 10^{-13} to 10^{-9} N μm^{-2} (refs 26, 27).

Being externally similar to MPIOs with comparable dipolar far-field decays, the structures can be spatially imaged via the same dephasing common to MPIOs; but in addition they can be differentiated spectrally and distinguished from spurious signal voids that confound SPIO/MPIO imaging. Depending on particle size, many different particle spectra can be acquired simultaneously from a single free induction decay following a broadband $\pi/2$ excitation. Alternatively, chemical shift imaging can spatially and spectrally resolve the tags simultaneously. Figure 3 demonstrates this spectral differentiation between individual particles. Because the spectra come from internal, rather than surrounding, water, spatial localization also improves substantially.

Diffusion-driven signal enhancement

Direct spectral imaging, however, is fundamentally limited by the relatively small number of nuclei within the structure that contribute to the signal. Our open structures, however, also allow an efficient analogue to magnetization transfer imaging[28,29], with diffusional exchange between water inside and outside the particle replacing traditional chemical exchange between bound and free protons. Therefore, using a preparatory set of $\pi/2$ pulses at the particle's shifted resonance to saturate out signal from a subsequent on-resonance pulse, the continual diffusion of fresh spins through the open particle structure can multiply its apparent signal volume. Scanned over off-resonant frequencies, this yields the so-called z-spectra[30] shown in Fig. 4b–e that also demonstrate how resonances can be engineered by manipulating structure geometry. Alternatively, fixing

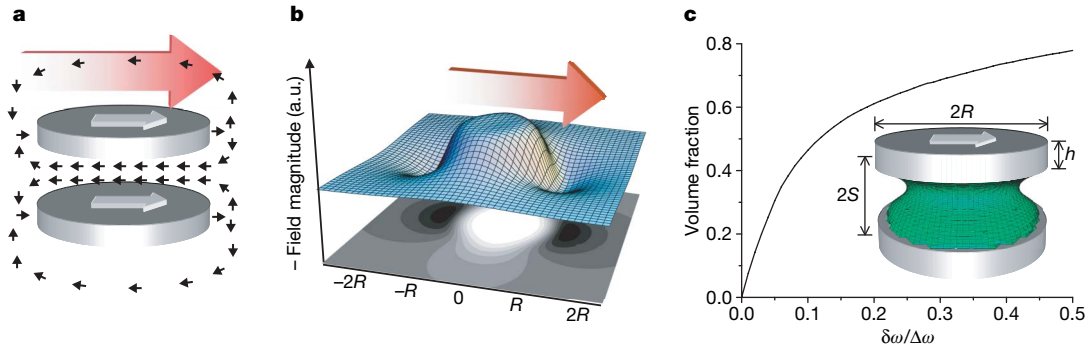


Figure 1 | Magnetic structure and field diagrams. a, Diagram of the field (black arrows) from two parallel disks magnetized to saturation by B_0 (red arrow). Non-magnetic spacer elements are omitted for clarity. **b**, Calculated (negative) field magnitude in the mid-plane through a typical magnetized disk set, contrasting its homogeneous nature between the disks with its rapid external decay. **c**, Calculated particle volume fraction that falls within a bandwidth, $\delta\omega$, about the particle's frequency shift, $\Delta\omega$. A sample numerical surface contour delineates the characteristic extent of this homogeneously shifted field region; all points inside the green contour shell have shifts within $\Delta\omega \pm \Delta\omega/50$.

the preparatory pulse train at the particle resonance allows spatial MRI of the transferred magnetization saturation, as shown in Fig. 5. By selectively blocking particle interiors, Fig. 5 also confirms that the signals arise specifically from water diffusing through the particles. Because the required time, τ_d, for self-diffusion to 'refresh' the internal water scales with R^2, the saturated magnetization falls only linearly with R, not with volume $\sim R^3$, as particle size is reduced. Without diffusion, the effective 'refresh' time would be limited by the longitudinal relaxation time, $T_1 \approx 2$–3 s. For water self-diffusivity, $D = 2.3 \times 10^{-9}\,m^2\,s^{-1}$, the distance diffused during this period, $(6DT_1)^{1/2} \approx 0.2$ mm, effectively sets the size below which open structures gain in sensitivity. This size is two orders of magnitude larger than typical micrometre-sized particles that might be used for cell labelling. Compared to structures that might have an enclosed internal volume of water, the gains in signal-to-noise ratio from diffusion through micrometre-sized open structures are therefore of order 10^4.

The double-disk structures afford a specific example of this magnetization exchange principle. Although we typically use first-principles Monte Carlo simulation (see Methods) to quantitatively predict exact diffusion-driven magnetization saturation levels, rough analytic approximation is also possible. Because of the high shifted-field homogeneity of the double-disk structures, we can suppress background signal while still saturating out about one-third of the volume between the disks via off-resonant excitation pulses with bandwidths just a few per cent of the particle's shift (Fig. 1c). For $h \ll 2S \approx R$, the magnetic moment of the water saturated in a single pulse is therefore $m_{pulse} \approx M_0 \pi R^3/3$, for M_0 the equilibrium B_0-aligned proton magnetization. Because not all the water exchanges between consecutive pulses, however, this per-pulse magnetic saturation falls with subsequent pulses. For an inter-pulse delay, $\tau_d = R^2/6D$, simulations show a resulting per-pulse average saturation of about $m_{pulse}/2$. The spatial distribution of any single pulse of this saturated magnetization at some later time, $t \gg \tau_d$, can be approximated by analogy to an instantaneous point-source diffusion problem, giving the local magnetization saturation $M_S(r,t) \approx (m_{pulse}/2)(4\pi Dt)^{-3/2}\exp(-r^2/4Dt)\exp(-t/T_1)$, where the final factor accounts for relaxation back into alignment with B_0, and r measures distance from the particle. Within a characteristic diffusion distance, $d \equiv (DT_1)^{1/2}$, a τ_d-spaced train of such pulses rapidly (order T_1) approaches the steady-state distribution, $M_S(r) \approx (M_0/4)(R/r)\exp(-r/d)$. Integrating over a (spherical) voxel of radius $R_v \gg R$, with $R_v \ll d$, gives the approximate fractional magnetization

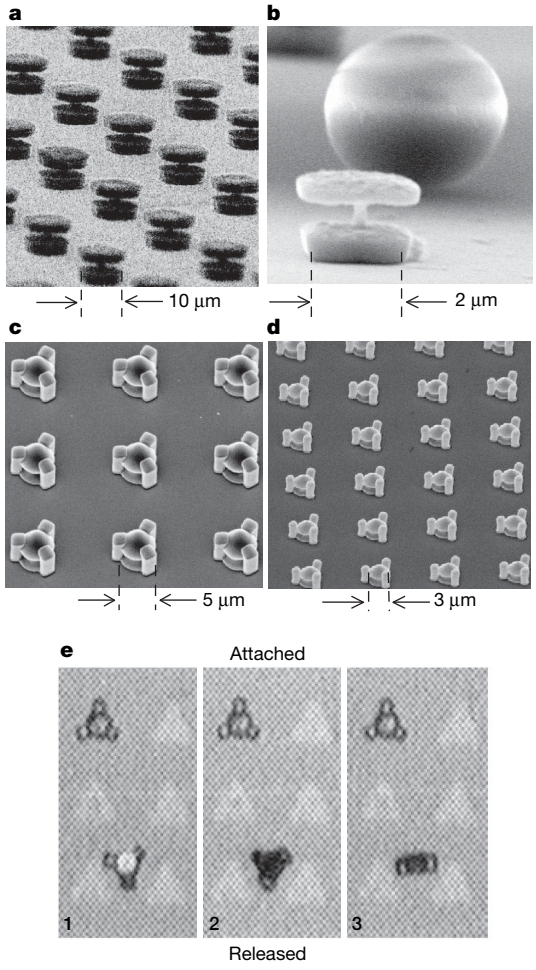


Figure 2 | Microfabricated magnetic structures. **a**, **b**, Scanning electron microscope (SEM) image of microfabricated double-disk magnetic structures with $R \approx 5$ μm (**a**) and $R \approx 1$ μm (**b**); the structures have non-magnetic internal supports. For relative size, a normal commercial 4.5 μm diameter MPIO (as commonly used for cell labelling/magnetic separation) is shown in the background in **b**. **c**, **d**, SEM image of externally supported double-disk structures with $R = 2.5$ μm (**c**) and R = 1.5 μm (**d**). In contrast to **a** and **b**, these particles demonstrate relatively thin magnetic layers, $h = 50$ nm, spaced $2S = 2$ μm (**c**) and 1 μm (**d**) apart. (The dome-like appearance of the top surfaces is due to a non-magnetic capping layer used during microfabrication.) These structures are robust, showing no discernible physical or magnetic change after month-long storage periods (both in and out of water). **e**, Optical micrograph contrasting a particle still attached to the substrate against an $R = 5$ μm particle released into water and automatically self-aligning with an applied magnetic field (of ~1 G) that is rotated from in-plane to out-of-plane in the sequence 1, 2, 3.

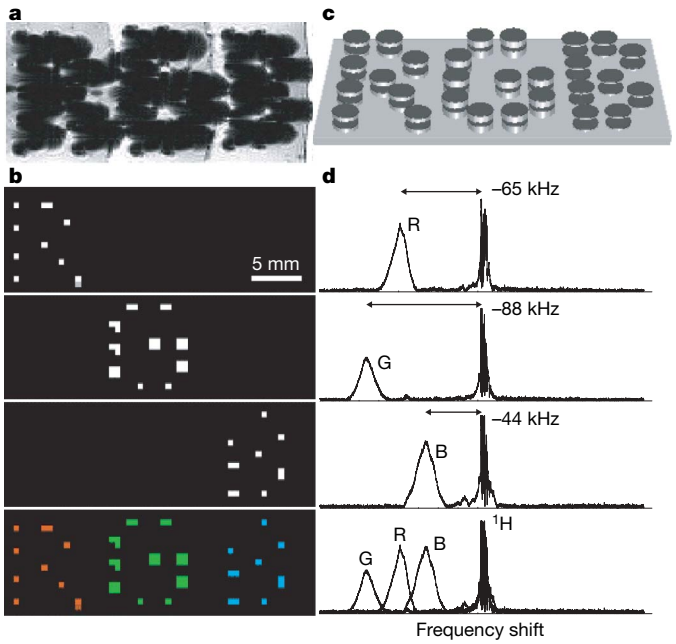


Figure 3 | Multi-spectral MRI. **a**–**d**, Chemical shift imaging of demonstration 1.25-mm-diameter particles magnetized by B_0. Particle frequency was varied by changing the thickness of electroplated nickel layers that formed the magnetizable disk pairs. As with normal SPIO detection, magnetic dephasing due to the particles' external fields enables the spatial imaging shown in the gradient-echo MRI (**a**). However, comparison between **a** and the chemical shift images (**b**) shows that the additional spectral information both differentiates between particle types and improves particle localization. The particles are shown schematically (not to scale) in **c**. With particle spectra (**d**, to the right of the corresponding chemical shift images in **b**) shifted well clear of the water proton line, different planes in the chemical shift imaging map isolate different particle types for unambiguous colour-coding with minimal background interference (**b**, bottom panel). (Although still visible in the gradient-echo image, the top corner particle of the letter 'B' was damaged, causing its shifted frequency peak to vanish.)

saturation of the water in the voxel immediately surrounding the particle as $M_S/M_0 \approx 0.3R/R_\gamma$. This linear rather than cubic scaling means, for example, that a sample $R = 2.5\,\mu m$ particle shown in Fig. 2c can saturate around 1–2% of a 50 μm radius voxel, even though its resonant field volume constitutes just 0.003% of that voxel. Such gains raise the possibility of simultaneous single microparticle imaging and spectral identification (as suggested in Fig. 5 legend) without the need for specialized microcoils[31]; indeed, all imaging described here was done with macroscopic surface and solenoidal coils up to several centimetres in diameter.

Comparison with traditional MRI agents

To compare the micro-engineered approach with traditional chemically synthesized agents, we turn from individual particle identification to detectable concentrations. Including continual longitudinal relaxation, the magnetic moment saturated out per particle pulsed over a period of $2T_1$ is $(m_{pulse}/2)(T_1/\tau_d)(1-e^{-2})$. Conservatively assuming at least 5% fractional saturation for reliable detection, required concentrations for micrometre-sized particles are therefore of order 10^{-14} M or, in elemental terms (assuming iron disks of aspect ratios similar to those of the particles in Fig. 2c), 0.01 mmol Fe l^{-1}. These concentrations are already below typical PARACEST concentrations used[7], an order of magnitude less than the clinical dosages of gadolinium relaxivity-based contrast agents in blood[5,32], and equal to those of SPIO agents[6]. However, as required concentrations scale with R^2, sub-micrometre structures that could be created using deep-ultraviolet or electron-beam lithography should substantially further reduce this concentration limit. Ultimately, the extent of the signal amplification that can be gained from this R^2 scaling is limited not by lithography, but by τ_d. By analogy with the 'slow-exchange' restriction[7] on chemical exchange processes, here diffusional exchange should not be so fast as to broaden the spectral peak by more than its shift. Fortunately, because large shifts can be generated, this exchange broadening becomes a limiting factor only below the 100 nm scale, at which point required metal concentrations would be in the nanomolar regime. Although this size scale may be regarded as a disadvantage over molecular-based agents, interest in MPIOs[19–21] indicates a growing range of applications for MRI contrast agents of similar, or even larger, size. Note, however, that while the 100 nm scale limits signal gained from further size reduction, it need not necessarily represent an absolute minimum structure size. Still smaller structures could be used by partially blocking access to the double-disk interior or by switching to an alternative less open structure to intentionally limit the effective exchange rate and keep τ_d within a desirable range. As sizes shrink further, the attendant shortening τ_d may also dictate that the preparatory RF pulse trains transform into quasi-continuous pulses; depending on the situation, such partial throttling of the water diffusion may also be desirable here.

Discussion

The faster imaging and increased safety margins that the structures' low concentration requirements imply are a consequence not only of faster allowable exchange rates, but also of the extended homogeneous field regions that can exchange many spins simultaneously, as

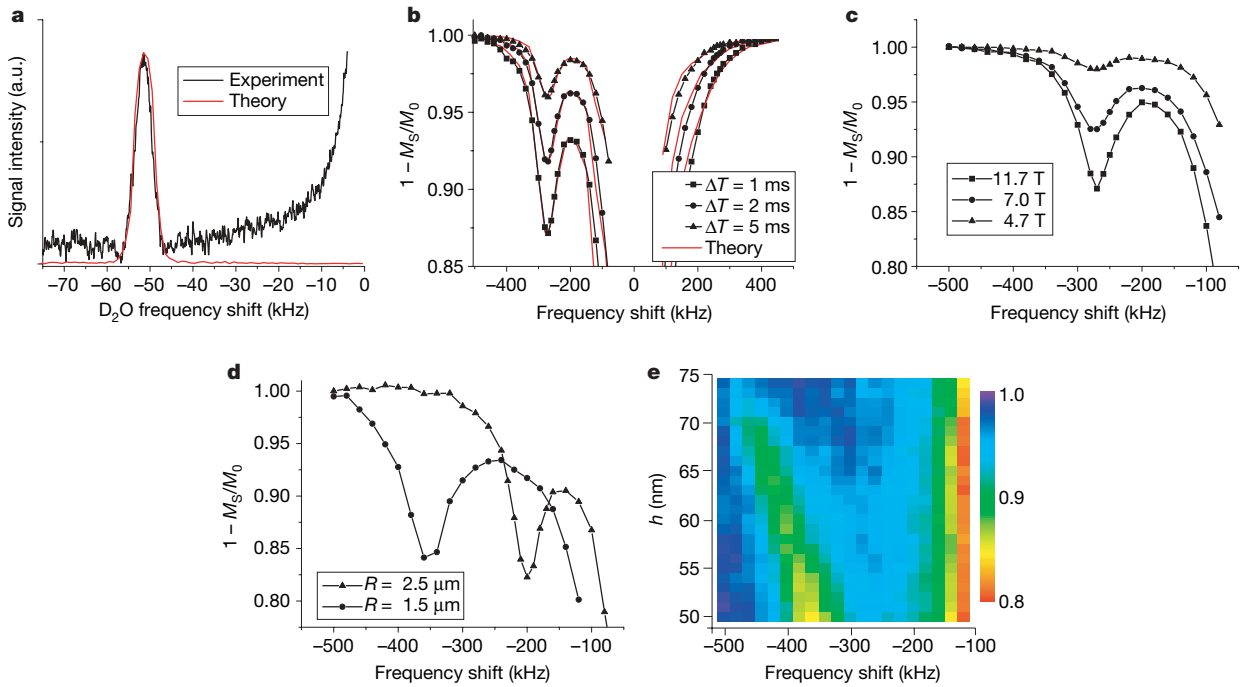


Figure 4 | Engineered spectral shifting. **a**, Fourier transformed spin-echo signal, showing direct imaging at 11.7 T of a spectrally shifted deuterium oxide peak from a set of $R = 12.5\,\mu m$ particles submerged in D_2O. Apart from overall signal magnitude, there are no free fitting parameters. **b**, $R = 2.5\,\mu m$ particle H_2O z-spectra taken at 7 T show increasing fractional saturation (M_S/M_0) with shortening delays, ΔT, between off-resonant $\pi/2$ pulses. Overlaid theory is derived from first-principles Monte Carlo simulation (see Methods) and contains no free fitting parameters. **c**, $R = 2.5\,\mu m$ particle H_2O z-spectra for $\Delta T = 2$ ms at three different field strengths, showing frequency shifting independent of B_0. **d**, H_2O z-spectra demonstrating different frequency shifts from structures with different values of R, but with fixed $h = 50$ nm and approximately constant $S/R \approx 0.3$–0.4. Because **c** and **d** assemble data from different MRI magnets and coils, comparative theory overlays are less meaningful, but data remain in agreement with theory. **e**, Continuous frequency-pulling engineered through continuously changing h (each row in the image shows the experimental H_2O z-spectrum for a different particle disk thickness, with the colour shading indicating the value of $1 - M_S/M_0$ at each point). For completeness, we show everywhere raw z-spectra of the shifted peaks atop the unshifted broadened water background; because the surrounding water broadening is approximately symmetric, however, this background can be eliminated by considering differences between corresponding positive- and negative-frequency saturations. All data are from first-generation test particle arrays with as yet still suboptimal geometries and ~10% interparticle frequency-shift variation due to cross-wafer manufacturing variation. Improved fabrication should reduce variation to below 1% and aid geometry optimization, substantially narrowing linewidths and increasing saturation levels.

opposed to the individual exchangeable proton sites of molecular complexes[7]. Micro-engineering also enables the use of biologically benign materials, allowing these field regions to be directly accessible and eliminating the efficiency-versus-toxicity trade-offs of agents based on chelated lanthanide ions[5,32]. Additionally, ferromagnetic or superparamagnetic materials ensure full saturation even for small B_0, enabling lower imaging fields while retaining large, field-independent shifts (Fig. 4c).

In principle, spectrally distinct, physiologically responsive indicators could also be formed by either encapsulating the particles, or filling their internal regions, to inhibit internal diffusion (Fig. 5) while leaving their external spatially trackable image-dephasings unaffected. If the material that blocks entry of water into the structures is chosen to be vulnerable to specific enzymatic attack, or to dissolution beyond a certain temperature or pH, subsequent water diffusion could effectively 'turn on' their spectral signals. Conversely, the spacer elements could be made from some dissolvable or reactive material to effectively modify or completely 'turn off' the spectral signals. Orientation-dependent sensors should also be possible by varying geometry to decrease magnetic self-alignment, yielding signals that appear or disappear depending on particle orientation. With spectral differentiation enabling multi-particle co-registration within the same voxel, a variety of multiplexed diagnostics can be envisaged. Additionally, their open structures and large shift ranges are well suited for flow and perfusion studies with multiple spin-labelled streams. Moreover, beyond MRI, their subcellular size may enable RFID-based microfluidics.

In conclusion, engineering local field environments over subcellular size-scales through tailored microstructures appears a promising avenue to a variety of new imaging and/or sensing mechanisms. Micrometre-sized structures can be microfabricated with a broad range of spectral coverage, and advanced lithographic techniques should enable substantial further decreases in the sizes of these structures, bringing them close to the sizes of nanoparticles at present in clinical use. Particularly encouraging are the design latitudes afforded by the high sensitivity of these micro-engineered agents, raising the possibility of a variety of additional microstructures that may similarly increase MRI functionality and impact.

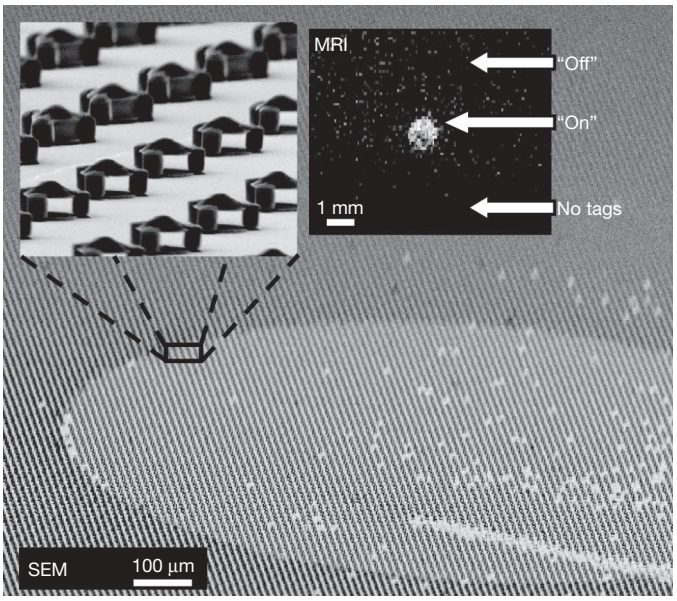


Figure 5 | Controlling diffusion to turn tags on or off. Main panel, high tilt-angle SEM image showing a square array of $R = 2.5$ μm particles. Except for a defined circular region, all particles have their interiors filled, blocking water diffusion. Top left inset, a higher magnification SEM image of the boundary between open and filled particles. Top right inset, the resulting background-subtracted chemical shift MRI showing transferred magnetization saturation from the particles' shifted resonance. Signal is visible from those particles that have water diffusing through their open interior region (labelled 'On') but not from those particles that have their interiors filled (labelled 'Off'). The bottom of the image shows a region that contains no particles (labelled 'No tags'), providing a null background signal comparison. A scratch (seen at the lower right corner) removed ~100 particles (about 10–20 per voxel). Its visibility in the magnetic resonance image suggests the potential for high-resolution imaging to spectrally distinguish individual such particles.

METHODS SUMMARY

Experimental set-up. Apart from the magnetic self-alignment experiments that involved freely floating particles in water, in order to enable more precise analysis, control experiments were performed on defined grids of test particles (13 mm × 13 mm square) attached to diced 15 mm × 15 mm Pyrex substrates on which the particles were originally microfabricated. Interparticle grid spacings (centre-to-centre) were typically 3 to 4 times the particle diameter, at which point numerical field calculations showed that any influence from the external fields of neighbouring particles had decayed to negligible levels. Individual Pyrex chips were sealed in custom-made holders filled with either water or deuterium oxide to a depth of at least 150 μm, sufficient to deeply submerge the particles and to continue well beyond the extent of any appreciable external particle field decays. Each of the water- or deuterium oxide-submerged samples were then individually placed next to, or inside, surface or solenoidal radio-frequency (RF) coils, respectively, for transmission/reception of the relevant NMR signals.

Numerical simulations. To help verify the physical understanding and analytical approximations presented, first-principles Monte Carlo simulations were also performed. These simulations modelled the effects of the applied RF field pulses and the (numerically calculated) fields of the magnetized double-disk microstructures on the local water, or deuterium oxide, nuclear spin evolution. Within the accuracy of our measurements we find good agreement with experiment (Fig. 4a, b), suggesting that the presented models capture the dominant physical processes involved.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 December 2007; accepted 21 April 2008.

1. Lauterbur, P. C. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature* **242,** 190–191 (1973).
2. Mansfield, P. & Grannell, P. K. NMR 'diffraction' in solids? *J. Phys. C* **6,** L422–L426 (1973).
3. Callaghan, P. T. *Principles of Nuclear Magnetic Resonance Microscopy* (Oxford Univ. Press, New York, 1991).
4. Nelson, K. L. & Runge, V. M. Basic principles of MR contrast. *Top. Magn. Reson. Imag.* **7,** 124–136 (1995).
5. Runge, V. M. & Wells, J. W. Update: Safety, new applications, new MR agents. *Top. Magn. Reson. Imag.* **7,** 181–195 (1995).
6. Weissleder, R. *et al.* Ultrasmall superparamagnetic iron oxide: Characterization of a new class of contrast agents for MR imaging. *Radiology* **175,** 489–493 (1990).
7. Woods, M., Woessner, D. E. & Sherry, A. D. Paramagnetic lanthanide complexes as PARACEST agents for medical imaging. *Chem. Soc. Rev.* **35,** 500–511 (2006).
8. Lanza, G. M. *et al.* $^{1}H/^{19}F$ magnetic resonance molecular imaging with perfluorocarbon nanoparticles. *Curr. Top. Dev. Bio.* **70,** 57–76 (2005).
9. Mason, W. T. (ed.) *Fluorescent and Luminescent Probes for Biological Activity* (Academic, London, 1999).
10. Bruchez, M. Jr, Moronne, M., Gin, P., Weiss, S. & Alivisatos, A. P. Semiconductor nanocrystals as fluorescent biological labels. *Science* **281,** 2013–2016 (1998).
11. Chan, W. C. W. & Nie, S. Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science* **281,** 2016–2018 (1998).
12. Alivisatos, P. The use of nanocrystals in biological detection. *Nature Biotechnol.* **22,** 47–52 (2004).
13. Elghanian, R., Storhoff, J. J., Mucic, R. C., Letsinger, R. L. & Mirkin, C. A. Selective colorimetric detection of polynucleotides based on the distance-dependent optical properties of gold nanoparticles. *Science* **277,** 1078–1081 (1997).
14. Haes, A. J. & Van Duyne, R. P. A nanoscale optical biosensor: Sensitivity and selectivity of an approach based on the localized surface plasmon resonance spectroscopy of triangular silver nanoparticles. *J. Am. Chem. Soc.* **124,** 10596–10604 (2002).
15. Nicewarner-Peña, S. R. *et al.* Submicrometer metallic barcodes. *Science* **294,** 137–141 (2001).
16. Dodd, S. J. *et al.* Detection of single mammalian cells by high-resolution magnetic resonance imaging. *Biophys. J.* **76,** 103–109 (1999).
17. Cunningham, C. H. *et al.* Positive contrast magnetic resonance imaging of cells labeled with magnetic nanoparticles. *Magn. Reson. Med.* **53,** 999–1005 (2005).

18. Bulte, J. W. M. *et al.* Magnetodendrimers allow endosomal magnetic labeling and *in vivo* tracking of stem cells. *Nature Biotechnol.* **19,** 1141–1147 (2001).
19. Hinds, K. A. *et al.* Highly efficient endosomal labeling of progenitor and stem cells with large magnetic particles allows magnetic resonance imaging of single cells. *Blood* **102,** 867–872 (2003).
20. Shapiro, E. M., Skrtic, S. & Koretsky, A. P. Sizing it up: Cellular MRI using micron-sized iron oxide particles. *Magn. Reson. Med.* **53,** 329–338 (2005).
21. Wu, Y. L. *et al.* In situ labeling of immune cells with iron oxide particles: An approach to detect organ rejection by cellular MRI. *Proc. Natl Acad. Sci. USA* **103,** 1852–1857 (2006).
22. Kotzar, G. *et al.* Evaluation of MEMS materials of construction for implantable medical devices. *Biomaterials* **23,** 2737–2750 (2002).
23. Voskerician, G. *et al.* Biocompatibility and biofouling of MEMS drug delivery devices. *Biomaterials* **24,** 1959–1967 (2003).
24. Chikazumi, S. *Physics of Ferromagnetism* (Oxford Univ. Press, New York, 1997).
25. Bozorth, R. M. *Ferromagnetism* (Van Nostrand, New York, 1951).
26. Sato, M., Wond, T. Z. & Allen, R. D. Rheological properties of living cytoplasm: Endoplasm of *Physarum* plasmodium. *J. Cell Biol.* **97,** 1089–1097 (1983).
27. Ashkin, A. & Dziedzic, J. M. Internal cell manipulation using infrared laser traps. *Proc. Natl Acad. Sci. USA* **86,** 7914–7918 (1989).
28. Henkelman, R. M., Stanisz, G. J. & Graham, S. J. Magnetization transfer in MRI: A review. *NMR Biomed.* **14,** 57–64 (2001).
29. Zurkiya, O. & Hu, X. Off-resonance saturation as a means of generating contrast with superparamagnetic nanoparticles. *Magn. Reson. Med.* **56,** 726–732 (2006).
30. Grad, J. & Bryant, R. G. Nuclear magnetic cross-relaxation spectroscopy. *J. Magn. Reson.* **90,** 1–8 (1990).
31. Olson, D. L., Peck, T. L., Webb, A. G., Magin, R. L. & Sweedler, J. V. High-resolution microcoil ^{1}H-NMR for mass-limited nanoliter-volume samples. *Science* **270,** 1967–1970 (1995).
32. Shellock, F. G. & Kanal, E. Safety of magnetic resonance imaging contrast agents. *J. Magn. Reson. Imag.* **10,** 477–484 (1999).

Acknowledgements We thank the Mouse Imaging Facility at the NIH for use of the 4.7T magnet, and A. Silva for use of the 7T magnet. This work was supported in part by the NINDS NIH Intramural Research Program. G.Z. also acknowledges support from a National Research Council fellowship award.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G. Z. (zabow@boulder.nist.gov).

METHODS

MRI experimental details. For the direct spectral detection experiment using water (spectra of Fig. 3), free induction decay (FID) signals following a spin-echo were acquired by sweeping through a range of frequencies covering the expected offsets produced by the particles. Shaped pulses with a gaussian profile were used to limit bandwidth spread into the bulk water peak (as compared to a hard pulse). Their bandwidths were, however, sufficient to cover the frequency profiles produced by the particles. Acquisitions for the spectra were 8,192 points in length, covering a bandwidth of 100 kHz. For the associated RGB image, three two-dimensional chemical shift images were acquired, covering the frequency ranges of the particle spectra. Images are integrations of the spectra over the different frequency ranges. In-plane resolution was 500×750 μm. Particle geometrical parameters were $R \approx 625$ μm, $2S \approx 500$ μm, and $h \approx 4$, 6 and 8 μm. Accidental impurities in the nickel disks of these structures led to a reduced $J_S \approx 0.4$ T. (All other structures had purer nickel with $J_S \approx 0.5$–0.6 T.)

For the direction detection experiment using D_2O (Fig. 4a), FIDs following a spin-echo were acquired using as large a bandwidth as our coil would allow, 50 kHz. Particle geometrical parameters were $R \approx 12.5$ μm, $2S \approx 10$ μm and $h \approx 0.5$ μm.

For the indirect detection experiments (Fig. 4b–e), the pulse sequence consisted of a series of off-resonance pulses (gaussian shape, 100 μs in length) for a period of a few T_1s, preceding an on-resonance 90° pulse for collection of an FID. Each point in the z-spectra represents the integral of this FID for a different off-resonance frequency of the preparatory pulse train. The gap between each pulse in the preparatory pulse trains was varied between 1 ms and 5 ms. For experiments at different field strengths (4.7, 7, 11.7 T), differing B_1 profiles from the different coils used may have led to some variations in the results. Particle geometrical parameters were $R \approx 2.5$ μm, $2S \approx 2$ μm and $h \approx 65$ nm for Fig. 4b, c; $R \approx 2.5$ μm, $2S \approx 2$ μm and $h \approx 50$ nm, and $R \approx 1.5$ μm, $2S \approx 1$ μm and $h \approx 50$ nm for Fig. 4d.

To demonstrate the spatial imaging using the indirect detection (Fig. 5), chemical shift images were acquired after a series of pulses at the predetermined offset frequency (in this case −330 kHz). A baseline image without the preparatory sequence was used to provide a subtraction image. The in-plane image resolution was 100×100 μm, with the thickness being determined by the 150 μm water depth. To speed up the imaging, the repetition time T_R was set to 500 ms, with the preparatory sequence being run continuously between each T_R. Particle geometrical parameters were $R \approx 2.5$ μm, $2S \approx 2$ μm and $h \approx 80$ nm.

It should be noted that all of the geometrical particle parameters listed represent approximate values only. Variation in particle parameters was in general dominated by slight variations in the exact purity of the nickel (and hence its precise magnetic saturation value) and by variations in the thickness of the nickel disk layers of about 10% throughout.

MRI numerical simulations. Simulations of the MRI experiments, results of which are seen in the theoretical curve fits to the data of Fig. 4a (direct spectral imaging) and Fig. 4b (indirect diffusional exchange based imaging), were derived from full first-principles Monte Carlo simulations purposely coded for analysing the double-disk structure experiments. To ensure accurate results, the Monte Carlo simulations simultaneously tracked the position, orientation and phase of upwards of several million simulated nuclear spins (with discrete time-steps down to a microsecond). These spins were modelled simultaneously randomly diffusing through a three-dimensional water volume (that matched the dimensions of the chip sample holder) surrounding a two-dimensional grid of double-disk structures that corresponded to the test chip layouts. Cyclic boundary conditions were used to reduce the number of double-disk structures that needed to be simulated. Larmor frequencies at each spatial location in this volume (used to compute the total accumulated phase of each spin over its random walk) were calculated based on numerically integrated calculations of the fields from the array of magnetized double-disk structures. 90° off- and on-resonance pulses (and for the direct detection, also 180° spin-echo pulses) were simulated via reorientation of only those spins that fell within the simulated resonant bandwidth of the applied pulses, as determined by the local Larmor frequency shifts at the location of each spin. Because self-diffusion distances over periods of order 100 μs (the typical pulse durations) can be appreciable at the micrometre scale, care was taken to simulate diffusion not just between applied RF pulses, but also during each RF pulse. Continual T_1-longitudinal relaxation was accounted for by reorientation of a set percentage of randomly chosen spins back into alignment with B_0 during each integration time-step. Signal acquisition was simulated based on the (time-varying) integrated field of all those spins within the readout bandwidth over the duration of the final simulated FID or spin-echo acquisition; since this integrated field included all magnetic field vector information (from orientation and phase of each spin), coherence/dephasing information was retained. Such coherence information affects the direct imaging spin-echo spectra, but, apart from loss of transverse coherence between RF 90° pulses, it does not affect the indirect diffusional exchange experiments. For the direct spectral imaging, the simulated acquired spin-echo signals were then numerically Fourier-transformed to give the final spectra (such as that shown in Fig. 4a); for the indirect diffusion-based imaging, the total percentage of spins saturated out, or essentially the integrated area under the simulated FID, gives the value of any point in the z-spectra shown in Fig. 4b (that is, the simulation is rerun for each point in the z-spectra instead of the single simulation run required for any direct imaging spectrum).

ARTICLES

The amphioxus genome and the evolution of the chordate karyotype

Nicholas H. Putnam[1,2], Thomas Butts[3], David E. K. Ferrier[4], Rebecca F. Furlong[3], Uffe Hellsten[1], Takeshi Kawashima[2]†, Marc Robinson-Rechavi[5,6], Eiichi Shoguchi[7]†, Astrid Terry[1], Jr-Kai Yu[8], Èlia Benito-Gutiérrez[9], Inna Dubchak[1], Jordi Garcia-Fernàndez[10], Jeremy J. Gibson-Brown[11]†, Igor V. Grigoriev[1], Amy C. Horton[11]†, Pieter J. de Jong[12], Jerzy Jurka[13], Vladimir V. Kapitonov[13], Yuji Kohara[14], Yoko Kuroki[15], Erika Lindquist[1], Susan Lucas[1], Kazutoyo Osoegawa[12], Len A. Pennacchio[1], Asaf A. Salamov[1], Yutaka Satou[7], Tatjana Sauka-Spengler[8], Jeremy Schmutz[16], Tadasu Shin-I[14], Atsushi Toyoda[15], Marianne Bronner-Fraser[8], Asao Fujiyama[15,17], Linda Z. Holland[18], Peter W. H. Holland[3], Nori Satoh[7]† & Daniel S. Rokhsar[1,2]

Lancelets ('amphioxus') are the modern survivors of an ancient chordate lineage, with a fossil record dating back to the Cambrian period. Here we describe the structure and gene content of the highly polymorphic ~520-megabase genome of the Florida lancelet *Branchiostoma floridae*, and analyse it in the context of chordate evolution. Whole-genome comparisons illuminate the murky relationships among the three chordate groups (tunicates, lancelets and vertebrates), and allow not only reconstruction of the gene complement of the last common chordate ancestor but also partial reconstruction of its genomic organization, as well as a description of two genome-wide duplications and subsequent reorganizations in the vertebrate lineage. These genome-scale events shaped the vertebrate genome and provided additional genetic variation for exploitation during vertebrate evolution.

Lancelets, or amphioxus, are small worm-like marine animals that spend most of their lives buried in the sea floor, filter-feeding through jawless, ciliated mouths. The vertebrate affinities of these modest creatures were first noted in the early part of the nineteenth century[1,2], and were further clarified by the embryologist Alexander Kowalevsky[3]. In particular, Kowalevsky observed that, unlike other invertebrates, amphioxus shares key anatomical and developmental features with vertebrates and tunicates (also known as urochordates). These include a hollow dorsal neural tube, a notochord, a perforated pharyngeal region, a segmented body musculature (embryologically derived from somites) and a post-anal tail. Together, the vertebrates, urochordates and lancelets (also known as cephalochordates) constitute the phylum Chordata, descended from a last common ancestor that lived perhaps 550 million years ago.

Although Kowalevsky, Darwin and others recognized the evolutionary relationship between chordate groups, the greater morphological, physiological and neural complexity of vertebrates posed a puzzle: how did the chordate ancestor—presumably a simple creature that resembled a modern amphioxus or ascidian larva—make such a transition?

Perhaps the most prevalent hypothesis for the origins of vertebrate complexity is founded on the ideas of Susumu Ohno (1970)[4], who proposed that vertebrate genomes were shaped by a series of ancient genome-wide duplications. In Ohno's original proposal, lancelet and vertebrates genomes were enlarged relative to the basic invertebrate complement by one or two rounds of genome doubling, although subsequent work suggested that these events occurred on the vertebrate stem after divergence of the lancelet lineage[5,6].

Although the sequencing of the human and other vertebrate genomes has shown that the gene number in vertebrates is comparable to, or only modestly greater than, that of invertebrates[7,8], evidence for large-scale segmental or whole-genome duplications on the vertebrate stem has mounted, with the parallel realization that most gene duplicates from such events are rapidly lost (reviewed in ref. 9). The relatively few surviving gene duplicates from the vertebrate stem provide evidence for ancient paralogous relationships between groups of human chromosomes[10–14] that plausibly arose from multiple rounds of whole-genome duplication before the emergence of modern vertebrates. However, the number, the timing and even the genomic scale of the duplication events, and their consequences for subsequent genome evolution, are poorly understood (for a review, see ref. 15), in part because the tunicate genomes are highly rearranged relative to the unduplicated early chordate karyotype (see below).

The Florida lancelet *B. floridae* (the generic name *Branchiostoma* refers to the characteristic perforated branchial arches) provides a critical point of reference for these studies[16]. This species and its

[1]Department of Energy Joint Genome Institute, Walnut Creek California 94598, USA. [2]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [3]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [4]The Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife KY16 8LB, UK. [5]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland. [6]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [7]Department of Zoology, Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan. [8]Division of Biology, California Institute of Technology, Pasadena, California 91125, USA. [9]National Institute for Medical Research, Mill Hill, London NW7 1AA, UK. [10]Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Avinguda Diagonal, 645, Barcelona 08028, Spain. [11]Department of Biology, Washington University in St Louis, St Louis, Missouri 63130, USA. [12]Children's Hospital of Oakland Research Institute, Oakland, California 94609, USA. [13]Genetic Information Research Institute, 1925 Landings Drive, Mountain View, California 94043, USA. [14]National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan. [15]RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [16]JGI Stanford Human Genome Center, 975 California Avenue, Palo Alto, California 94304, USA. [17]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan. [18]Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093-0202, USA. †Present addresses: Okinawa Institute of Science and Technology (OIST), Uruma, Okinawa 904-2234, Japan (T.K., E.S. and N.S.); Genome Sequencing Center, Department of Genetics, Washington University in St Louis School of Medicine, St Louis, Missouri 63108, USA (A.C.H.); Institute for Evolutionary Discovery, 909 Hiawatha Drive, Mount Pleasant, Michigan 48858, USA (J.J.G-B.).

relatives (collectively also known as amphioxus, derived from the Greek *amphi+oxys*, 'sharp at both ends') are widely regarded as living proxies for the chordate ancestor, in part owing to the general similarity of the modern amphioxus to putative fossil chordates from the early Cambrian Chengjiang fauna (*Yunnanozoon lividum*[17], and the similar *Haikouella lanceolatum*[18]) and the middle Cambrian Burgess Shale (*Pikaia gracilens*[19]), although controversy remains (see, for example, refs 20,21). The study of key developmental genes in amphioxus has shed light on the evolution of such vertebrate organs as the brain, kidney, pancreas and pituitary, and of the genetic mechanisms of early embryonic patterning in general (reviewed by refs 22–24). Amphioxus has also served as a genomic surrogate for the protovertebrate ancestor in studies of the Hox cluster[25], in studies of specific genomic regions[26–28], and as an outgroup in numerous gene family studies (reviewed in ref. 29).

Here we report the draft genome sequence of the Florida lancelet and compare its structure with the genomes of other animals. Robust phylogenetic analysis of gene sequences and exon–intron structures confirms recent proposals that tunicates are the sister group to vertebrates, with lancelets as the most basal chordate subphylum, and that the combined echinoderm–hemichordate clade is sister to chordates. Through a comparative analysis, we identify 17 ancestral chordate linkage groups that are conserved in the modern amphioxus and vertebrate genomes despite over half a billion years of independent evolution. Over 90% of the human genome is encompassed within these linkage groups, which display a tell-tale fourfold redundancy that is consistent with whole-genome quadruplication on the vertebrate stem. Comparison with sequences from the sea squirt, lamprey, elephant shark and several bony fish constrains the timing of the whole-genome events to after the divergence of vertebrates from tunicates and lancelets, but before the split between cartilaginous and bony vertebrates. Within the resolution of our analysis, we find evidence for rounds of genome duplication both before and after the split between jawless vertebrates (for example, lamprey) and jawed vertebrates, although a period of octoploidy encompassing the divergence of jawless and jawed vertebrates remains a possibility. Although most duplicate genes from these whole-genome events have been lost, a disproportionate number of genes involved in developmental processes are retained.

Genome sequence

We sequenced the ~520 megabase (Mb) amphioxus genome using a whole-genome shotgun strategy[30] from approximately 11.5-fold redundant paired-end sequence coverage produced from random-sheared libraries with a range of insert sizes (Supplementary Note 2). Genomic DNA was prepared from the gonads of a single gravid male collected from Tampa Bay, Florida in July 2003, and exhibited extensive allelic variation (3.7% single nucleotide polymorphism, plus 6.8% polymorphic insertion/deletion; Supplementary Note 4). This is the highest level of sequence variation reported in any individual organism, exceeding that found in the purple sea urchin[31]. Assembly version 1 reports both haplotypes separately, whereas in assembly version 2 a single haplotype is selected at each locus (Supplementary Fig. 63). Assembly version 2 spans 522 Mb, with half of this sequence in 62 scaffolds longer than 2.6 Mb.

Currently there are no physical or genetic maps of amphioxus, so we could not reconstruct the genome as its 19 pairs of chromosomes[32]. Nevertheless, because half of the predicted genes are contained in scaffolds containing 138 or more genes, the current draft assembly is sufficiently long-range to permit useful analysis of conserved synteny with other species, as shown below. Comparison of the assembled sequence with open reading frames derived from expressed sequence tags (ESTs, see below) shows that the assembly captures more than 95% of the known protein-coding content, and comparison to finished clone sequences demonstrates the base-level and long-range accuracy of the assembly (Supplementary Note 2).

Protein-coding genes and transposable elements

We estimate that the haploid amphioxus genome contains 21,900 protein-coding loci. This gene complement was modelled with standard methods tuned for amphioxus, integrating homology and *ab initio* gene prediction methods with more than 480,000 ESTs derived from a variety of developmental stages[24] (Supplementary Note 3). Approximately two-thirds of the protein-coding loci (15,123) are captured in both haplotypes. Transposable elements constitute ~30% of the amphioxus genome assembly (Supplementary Table 5) and belong to >500 families. On the basis of their bulk contribution to the genome size, DNA transposons are twice more abundant than retrotransposons.

Polymorphism

The distribution of observed local heterozygosity over short length scales obeys a geometric distribution (Supplementary Fig. 1), consistent with the prediction of the random mating model, as observed in *Ciona savignyi*[33], with a population mutation rate $4\mu N_e = 0.0562$, where μ is the per generation mutation rate and N_e is the effective population size. High heterozygosity can, in principle, be explained by: (1) a large effective population size maintained over many generations, (2) a high mutation rate per generation, or (3) the recent mixing of previously isolated populations; in the latter case, a geometric distribution of local heterozygosity would not be expected. Assuming a typical metazoan mutation rate on the order of one to ten substitutions per gigabase (Gb) per generation[34,35], this observed heterozygosity between alleles suggests a large but plausible effective breeding population on the order of millions of individuals. The observed heterozygosity shows correlations at short distances that decay on scales greater than ~1 kb, indicating extensive recombination in the population (Supplementary Fig. 2). An analysis of the ratio K_a/K_s of non-synonymous to synonymous substitutions shows evidence of purifying selection comparable to that found between mammalian species (Supplementary Note 4). Insertion/deletion polymorphisms are common, as found in other intra- and inter-genome comparisons[36] (Supplementary Fig. 3). Structural variation between haplotypes also includes local inversions and tandem duplications (Supplementary Fig. 63).

Deuterostome relationships

With the draft amphioxus sequence in hand, we reconsidered the relationships within chordates and between deuterostome phyla (chordates, echinoderms and hemichordates). The traditional placement of lancelets as sister to vertebrates, with tunicates as the earliest diverging chordate subphylum, has recently been questioned[37,38]. A preliminary study[39] using 146 gene loci (33,600 aligned amino acid positions) and trace data from the present amphioxus genome project found support for tunicates as the sister to vertebrates. This analysis, however, also suggested (albeit with limited statistical support) that amphioxus is more closely related to echinoderms than to tunicates or vertebrates, which would render chordates a paraphyletic group. A second study with a similarly sized set of genes but more diverse deuterostome taxa supported the early branching of the cephalochordate lineage, but not the close relationship of amphioxus and echinoderms[40].

To address the controversial phylogenetic position of amphioxus, we analysed a much larger set of 1,090 orthologous genes (see Supplementary Note 5). Both bayesian and maximum likelihood methods support the new chordate phylogeny[38–40] in which cephalochordates represent the most basal extant chordate lineage, with tunicates (represented by both *Ciona intestinalis* and *Oikopleura dioica* in our analysis) sister to vertebrates but with long branches that indicate higher levels of amino acid substitution (Fig. 1). Individual gene trees also lend support to this topology; genes supporting tunicates as a sister group to vertebrates outnumber those with amphioxus in this position by a 2:1 ratio. An analysis of intron gain and loss in deuterostomes provides independent support for amphioxus as the basal extant chordate subphylum (see below).

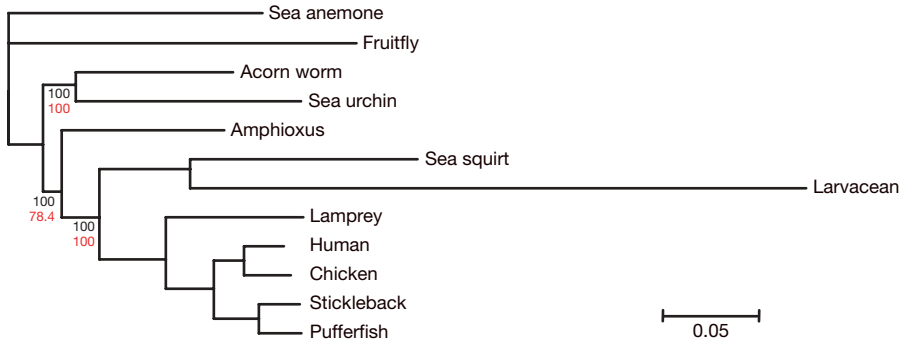


Figure 1 | Deuterostome phylogeny. Bayesian phylogenetic tree of deuterostome relationships with branch length proportional to the number of expected substitutions per amino acid position, using a concatenated alignment of 1,090 genes. The scale bar represents 0.05 expected substitutions per site in the aligned regions. Long branches for sea squirt and larvacean indicate high levels of amino acid substitution. This tree topology was observed in 100% of sampled trees (see Supplementary Note 5). Numbers in red indicate bootstrap support under maximum likelihood. Unlabelled nodes were constrained.

We group echinoderms (that is, the purple sea urchin) and hemichordates (that is, the acorn worm) together (ambulacrarians) as sister to a monophyletic chordate clade, as in ref. 41 but in contrast to the suggestion of ref. 39. With the exception of the long-branched tunicates, the maximum likelihood tree suggests a roughly constant evolutionary rate of peptide change across the deuterostome tree, although an excess of substitutions is found in the vertebrates relative to the predictions of a simple molecular clock model.

Intron evolution

To assess the evolution of gene structure within the deuterostomes and chordates, we compared the position and phase of amphioxus introns to those in other animals. Amphioxus and human (along with other vertebrates) share a large fraction of their introns (85% in alignable regions), which match precisely in both position and phase (Supplementary Note 6), as was also found in the sea anemone *Nematostella vectensis*[42]. We found that the intron-rich gene structures of the eumetazoan ancestor were carried forward to the common chordate ancestor with relatively few gains or losses. The tunicates *C. intestinalis* and *Oikopleura dioica*, however, share many fewer introns with vertebrates[43] or amphioxus.

Notably, intron presence or absence carries a significant (as measured by bootstrap values) phylogenetic signal, and bayesian analysis of the associated character matrix supports the sister relationship between tunicates and vertebrates (Supplementary Fig. 8; see Methods). This is evidently due to shared gain or loss of introns along the stem group leading to their common ancestor, which remarkably is still detectable despite additional extensive secondary losses, and modest gains, in the tunicate lineages. Thus, intron dynamics provide independent support for the new chordate phylogeny.

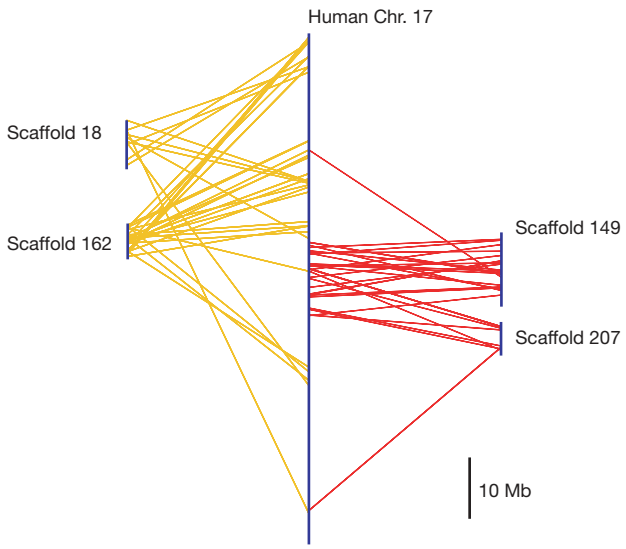


Figure 2 | Amphioxus–human synteny. Four amphioxus scaffolds from the non-redundant version 2 assembly with synteny to human chromosome (Chr.) 17. Note that orthologous genes from these scaffolds are concentrated in specific regions of the chromosome, and that several scaffolds (for example, 18 and 162, or 149 and 207) have a high density of hits to the same segments of the chromosome, which enables a partitioning of the human genome into 135 ancient segments. Supplementary File 5 contains an Oxford grid tabulating the number of orthologues for each scaffold-segment pair.

Chordate gene families and novelties

Through comparison of the amphioxus gene set with those of other animals, we identified 8,437 chordate gene families with members in amphioxus and other chordates that each nominally represent the modern descendants of a single gene in the last common chordate ancestor (Supplementary Note 7). That ancestor certainly possessed more genes than this number, but the others are inaccessible to us now owing to subsequent sequence divergence and/or gene loss in the living chordates. Through subsequent gene family expansions (by means of both local and/or genome-wide duplications), these families account for 13,610 amphioxus genes, 13,401 human genes and 7,216 *C. intestinalis* genes. The markedly lower number of descendant genes in *C. intestinalis* is largely due to gene loss[44], with the present analysis identifying 2,251 ancient chordate genes missing in this genome sequence. We found 8 apparent chordate stem gene losses (that is, genes found in sea urchin and at least one of fly and sea anemone, but not in vertebrates, amphioxus or *C. intestinalis*). A list of these genes can be found in Supplementary Table 10.

We identified 239 apparent chordate gene novelties, that is, gene families represented in amphioxus and at least one vertebrate or *Ciona*, but without an obvious direct counterpart in non-chordate genomes. These can be characterized[42] as 137 families with no detectable sequence similarity to non-chordate genes (type I novelties), 10 containing one or more chordate-specific domains linked to pre-existing metazoan domains (type II novelties), and 92 with chordate-specific combinations of pre-existing metazoan domains (type III novelties; see Supplementary Note 7). These gene families and others of special interest to vertebrate biology are discussed in a separate paper[45].

Amphioxus–vertebrate synteny

We have found extensive conservation of gene linkage on the scale of whole chromosomes (macro-synteny) between the amphioxus genome and those of vertebrates (represented in our analysis by human, chicken and teleost fish), but only limited conservation of local gene order (micro-synteny). Through comparative analysis of these conserved features, we reconstructed the gene complements of 17 linkage

groups (that is, proto-chromosomes) of the last common chordate ancestor. When vertebrate genomes are analysed in the light of these putative ancestral chordate chromosomes, a clear pattern of global fourfold conserved macro-synteny is found, demonstrating that two rounds of whole-genome duplication occurred on the vertebrate stem.

Reconstruction of chordate linkage groups

To identify ancestral chordate linkage groups, we first noted that many individual amphioxus scaffolds show conserved syntenic associations with human chromosomes, reflecting conserved linkage between the two genomes (Fig. 2; see also Oxford Grid in Supplementary File 5; for simplicity, we emphasize the amphioxus–human comparison in the main text, and include similar results for chicken, stickleback and pufferfish as Supplementary Information). Ninety-six scaffolds (out of 129 that possess 20 or more independent vertebrate orthologues) have a significant ($P < 0.05$, multiple test corrected) concentration of orthologues on one or more human chromosome. In contrast, only 12 *C. intestinalis* scaffolds (out of 134 that contain 20 or more vertebrate orthologues) show significant synteny to human chromosomes.

Genes on individual amphioxus scaffolds have orthologues that are generally concentrated in specific regions of vertebrate chromosomes (Fig. 2). Furthermore, multiple amphioxus scaffolds typically exhibit hits to the same sets of human chromosomal regions. Within each region, only limited conservation of gene order is observed (Methods; Supplementary Note 8). This pattern of conserved synteny shows that genome rearrangements have not erased the imprint of the genome organization of the last common chordate ancestor from the present human and amphioxus genomes. By using this pattern, we identified 135 human chromosomal segments (listed in Supplementary Table 14) that retain relict signals of the ancestral chordate karyotype despite chromosomal rearrangements in each lineage (Methods). These segments span a mean of 170 genes.

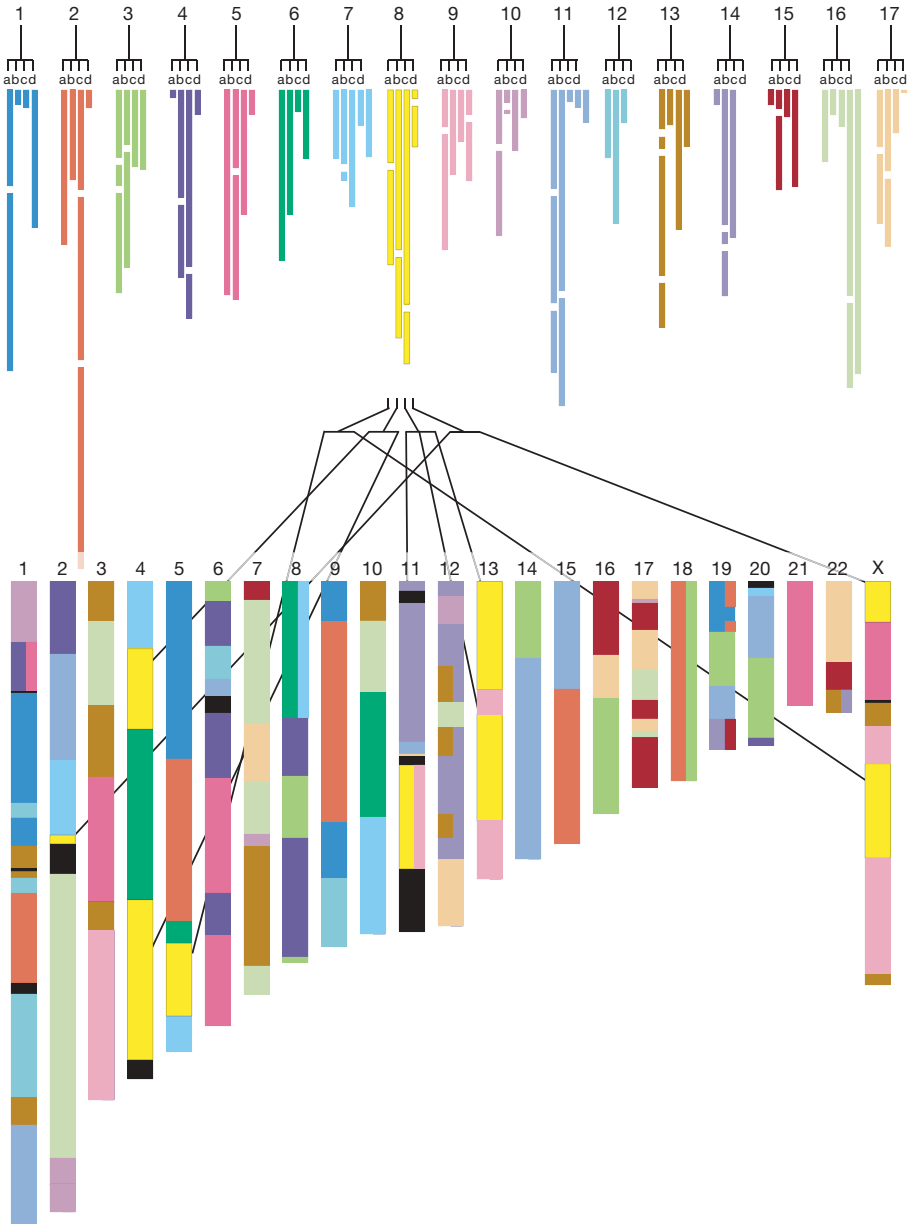


Figure 3 | Quadruple conserved synteny. Partitioning of the human chromosomes into segments with defined patterns of conserved synteny to amphioxus (*B. floridae*) scaffolds. Numbers 1–17 at the top represent the 17 reconstructed ancestral chordate linkage groups, and letters a–d represent the four products resulting from two rounds of genome duplication. Coloured bars are segments of the human genome, shown grouped by ancestral linkage group (above), and in context of the human chromosomes (below).

We exploited the pattern of amphioxus–human synteny to identify 17 ancient chordate linkage groups (CLGs) by clustering both amphioxus scaffolds and human chromosomal segments according to their pattern of hits in the other genome (Methods). The resulting 'dot plot' (Supplementary Fig. 64) shows that orthologues are concentrated in 17 distinct blocks. Within each block, gene order is considerably scrambled. The natural interpretation of these blocks is that each represents an ancient chordate linkage group that evolved into a defined group of chromosomal segments in amphioxus, human, chicken and teleost fish.

We tested our interpretation of the chordate linkage groups as coherently evolving segments by using fluorescent *in situ* hybridization (FISH) to demonstrate that 15 out of 16 scaffolds from CLG 15 localized to a single amphioxus chromosome. (FISH of the BACs corresponding to the sixteenth scaffold were ambiguous; see Supplementary Table 2.) Similarly, an independent study of amphioxus cosmids containing NK group homeobox genes in CLG 7 found that they localize to several distant regions of a single chromosome in amphioxus[27,28]. These data support the claim that the 17 putative ancestral chordate linkage groups have been maintained in modern amphioxus as coherent chromosomal segments. The 19 pairs of modern amphioxus chromosomes, however, imply at least (see below) two subsequent fissions in the amphioxus lineage.

Within these segments, nearly 60% of the human genes that possess amphioxus orthologues participate in the conserved linkage groups. This represents a lower bound, because short amphioxus scaffolds are less likely to be assigned to CLGs. Conversely, 88% of amphioxus gene models on scaffolds assigned to a CLG have their human orthologue in a conserved position (that is, in the same CLG). Remarkably, to the resolution of our analysis, some entire chromosomes (for example, human 18 and 21; chicken 7, 12, 15, 19, 21, 24 and 27) and chromosome arms (including human 5p) seem to have maintained their integrity (with local scrambling and some gene gain and loss) since the last common chordate ancestor. The CLGs defined by comparing amphioxus and vertebrate genomes also provide a new perspective on tunicate genome evolution, because it appears that *C. intestinalis* chromosomes 10, 12 and 14 are each relicts of a single CLG (11, 5 and 8, respectively), and other conserved linkages are evident (Supplementary Fig. 14).

Quadruple conserved synteny

We can trace the evolution of chordate genomes through time using two additional types of evidence. First, we can constrain the timing of specific chromosome breaks by parsimony analysis of conserved synteny across human, chicken and teleost genomes. Second, we can use the presence (or absence) of paralogous gene pairs to identify segments derived from the same chordate proto-chromosome by duplication (or fission).

For example, five groups of human segments from chromosomes 1, 5, 9 and 19 cluster together in CLG 1 (Supplementary Note 8; Supplementary Fig. 64). The segment pair 1.5/7 from chromosome 1 is related to each of the others by a significant concentration of ancient gene paralogues (17 to 31 pairs, $P < 1 \times 10^{-10}$), indicating that it is related to the others by duplication. In contrast, only a single pair of ancient paralogues relates segments 5.1 and 9.1/3, and orthologues of the genes in these segments occur predominantly on the same chromosomes of both pufferfish and stickleback. Thus, 5.1 and 9.1/3 were probably created by breakage of a single ancestral segment of the genome of the bony vertebrate ancestor. If 5.1 and 9.1/3 are virtually merged, then all remaining pairings of human segments from CLG 1 show a significant excess of ancient paralogues, consistent with amplification to four through two successive duplications.

To obtain a genome-wide view of the history of chromosomal evolution on the vertebrate stem, we applied a similar analysis systematically to the 17 CLGs by exhaustively evaluating all partitionings of human genome segments, and using a parsimony criterion to identify the most likely reconstruction. The most parsimonious partitionings of human segments into paralogy groups is summarized in Supplementary Table 1, and is diagrammed in Fig. 3. This analysis shows that most of the human genome (112 segments spanning 2.68 Gb, or 95% of the euchromatic genome) was affected by large-scale duplication events on the vertebrate stem before the bony vertebrate radiation (that is, the teleost/tetrapod split), and that nearly all of the ancient chordate chromosomes were quadruplicated (Supplementary Fig. 9).

This pattern of genome-wide quadruple conserved synteny[15] definitively demonstrates the occurrence of two rounds of whole-genome duplication (2*R*) and provides a comprehensive reconstruction of the evolutionary origin of the human chromosomes (and those of other jawed vertebrates) through these duplications on the vertebrate stem. This characterization extends previous lines of evidence for whole-genome duplication events based on comparative studies of specific regions of interest across chordate genomes (for example, the Hox cluster[25] and the major histocompatibility complex region[28,46]) and the analysis of vertebrate gene families (reviewed in ref. 29), as well as the identification of paralogous segments and chromosomal relationships within the human genome[10,13,14,47]. A manual, phylogeny-based analysis of the four scaffolds making up the NK homeobox-containing paralogon was in agreement with these results (Methods).

Timing of events on the vertebrate stem

The amphioxus–human synteny analysis presented here demonstrates that two rounds of whole-genome duplication occurred on the vertebrate stem after the divergence of cephalochordates but before the split of teleosts and tetrapods. The next question is whether these two genome-scale duplications happened in rapid succession or even effectively simultaneously, or were separated in time[15]. We sought to resolve the 2*R* events relative to the divergence of cartilaginous fish, urochordates and jawless vertebrates (for example, lamprey).

Sample sequencing from the elephant shark *Callorhynchus milii*, for example, demonstrates significant conserved macrosynteny between cartilaginous fish and humans, because pairs of genes that are ~35–40 kb apart in the elephant shark genome are also linked on the human genome[48]. These links occur predominantly within the human segments defined above, indicating that the orthologous chromosome segments are also found in the elephant shark genome (Supplementary Note 9). Furthermore, previous analysis of phylogenetic topologies dated all duplications before the split between cartilaginous and bony vertebrates[49]. Therefore, 2*R* occurred before this split[48]. Similarly, the preservation of several CLGs as intact single chromosomes in *C. intestinalis* (Supplementary Fig. 14) implies that both rounds of duplication occurred after the divergence of the urochordate lineage.

Sequencing of the repeat-rich lamprey genome has not generated enough long scaffolds to permit large-scale analysis of synteny[50]. To infer the timing of 2*R* relative to the divergence of the lamprey lineage, we generated a set of ~50,000 ESTs from the sea lamprey *Petromyzon marinus* (Supplementary Note 3) and analysed the phylogenetic topology of 358 gene families that include pairs of synteny-confirmed human paralogues produced during 2*R*. The results are summarized

Table 1 | Timing of whole-genome duplications

Gene '*X*' from	Number of phylogenies attempted	Number of resolved trees	Number of resolved trees with '*X*' as ingroup	Percentage (mean ± standard error)
Ciona intestinalis (sea squirt)	736	273	34	12 ± 2%
Petromyzon marinus (lamprey)	358	159	93	58 ± 6%
Fugu rubripes (pufferfish)	1009	389	351	90 ± 5%

in Table 1, along with a parallel analysis using sea squirt and pufferfish for comparison. We find that ~58% of the resolved four-gene phylogenies place the lamprey gene closer to one of the human paralogues; this is similar to the results of ref. 51 but analyses a tenfold larger set of gene families (Supplementary Fig. 2). This result is clearly distinct from that expected if the lamprey lineage diverged either much before (as for sea squirt) or after (as for pufferfish) 2*R*. The remaining scenarios are either that the jawed vertebrate and lamprey lineages diverged in the period between two well-separated whole-genome duplications, or that one, or both, of the 2*R* whole-genome events occurred nearly coincident with the lamprey lineage divergence. The time interval that distinguishes 'nearly coincident' from 'well-separated' is determined by the process of rediploidization, during which most gene duplicates are lost and the sequences of the surviving paralogues diverge[9,15].

Karyotypic changes in the vertebrate and tunicate lineages

From the 17 ancestral chordate linkage groups, 2*R* nominally produced $17 \times 4 = 68$ proto-vertebrate segments, although this naive inference assumes that, first, all duplicated segments were retained and, second, no fusions, fissions or additional segmental duplications occurred during 2*R*. Some 2*R*-produced segments, however, are consistently linked in contemporary bony vertebrate genomes (for example, 12b and 1d, which co-occur on human chromosome 1, chicken chromosome 8 and stickleback linkage groups III and VIII), indicating a fusion before the bony vertebrate (osteichthyan) ancestor. We found evidence for at least 20 such fusions (Supplementary Note 8). Allowing for a range of nearly parsimonious reconstructions of 2*R*, we estimate that the bony vertebrate ancestor had between 37 and 49 chromosomes. Additional fusions on the teleost stem reduced this number to 12 (refs 52–56) before the teleost-specific genome duplication. On the tetrapod stem, the chicken and human genomes share 4 fusions of bony vertebrate segments, suggesting 33–45 chromosomes. These are consistent with recent estimates based on intra-vertebrate comparisons[47,52–56].

Ancient developmental gene linkages

The amphioxus genome has also retained ancient local gene linkages (micro-synteny) in addition to conserved macro-synteny. In some cases, local linkages are even older than the chordates, and date back to the bilaterian ancestor or earlier. As an example, we considered gene families that expanded by tandem duplication early in animal evolution, specifically, the Antennapedia (ANTP) and Paired (PRD) classes of homeobox genes and the *Wnt* gene family. We examined how frequently these genes are still neighbours in the amphioxus genome, and discovered five new examples of ancient pairs or clusters of ANTP or PRD genes: (1) *Otx* and *goosecoid*, (2) *Mnx* and *ro*, (3) *Nkx2-1* and *Nkx2-2*, (4) *Nkx6*, *Nkx7*, *Lbx* and *Tlx*, and (5) *En*, *Nedxa*, *Nedxb* and *Dll* (Supplementary Table 3). These gene pairs or clusters (along with the well-known *Hox*, *ParaHox* and *NK* linkages) originated by tandem duplication before the divergence of bilaterians, yet their tight linkage has not been disrupted by genome rearrangement. The *Nkx2-1*/*Nkx2-2* gene pair has been retained in vertebrates (and is duplicated), but in every other example the tight linkage (clustering) has been lost in the human genome. None of the five newly described examples are retained in the *Drosophila melanogaster* genome. The situation in the *Wnt* gene family is a little different, because both amphioxus and *Drosophila* have retained tight linkages that have been disrupted in the human lineage owing to genome duplication followed by differential gene loss (Fig. 4). These results underscore the fact that the amphioxus genome has undergone less genomic rearrangement than the human and other vertebrate genomes since their shared ancestor more than half a billion years ago.

Impact of whole-genome duplications

How many duplicate genes survive in modern vertebrate genomes from the two genome-wide events? Twenty-five per cent (2,131) of the ancestral chordate gene families (out of the 8,437 indicated above) have two or more ancient vertebrate paralogues ('ohnologues') that were evidently produced by ancient gene duplication(s) after the divergence of amphioxus (see also Supplementary Fig. 3). Of these, 1,489 (70%) are embedded within paralogous segments from our reconstruction of 2*R*, as portrayed in Fig. 3, and were plausibly created through 2*R*. These retention rates for 2*R*-duplicated genes are comparable to other estimates based on large-scale gene phylogenies[10,14,57,58]. Similar retention rates were found for the ~350-million-year-old teleost-fish-specific duplication[55,59–61] and were estimated for the ~40-million-year-old genome duplication found in the frog *Xenopus laevis*[62].

Gene duplicates from 2*R* that have been retained in modern genomes are significantly enriched for functions associated with signal transduction, transcriptional regulation, neuronal activities and developmental processes (Supplementary Table 18, Methods). For example, genes implicated in signal transduction are more than twice as likely to be retained in two or more copies from 2*R* compared to the overall retention rate. These results are consistent with the hypothesis that paralogues created by whole-genome duplication were recruited for roles in the development of novel features of vertebrate biology, and with similar biased retention in teleost fishes[61]. Whole-genome duplications, however, may have allowed entire molecular pathways to be duplicated and sub-functionalized coincidently (reviewed in ref. 63). Whereas similar numbers of gene duplicates are found in amphioxus relative to the chordate ancestor, different gene classes have been expanded, and the mechanism of gene duplication is different (Supplementary Note 7).

Conserved non-coding sequences

Inspired by the extensive conserved synteny between amphioxus and vertebrates, we searched for conserved non-coding sequences that might reflect ancient chordate *cis*-regulatory elements. Genome-scale comparisons between mammals and teleost fish have revealed up to 3,100 conserved non-coding sequences, most of which function as tissue-specific enhancers[64,65]. At greater phylogenetic distances, no conservation outside of coding sequences and conserved microRNAs[66] has been identified so far. By aligning the amphioxus and human genomes (Supplementary Note 10), 77 putative chordate conserved non-coding elements were identified (>60% identity over >50 bp), after excluding transcribed or repetitive sequences and requiring conservation in at least one other vertebrate. Of these, 16

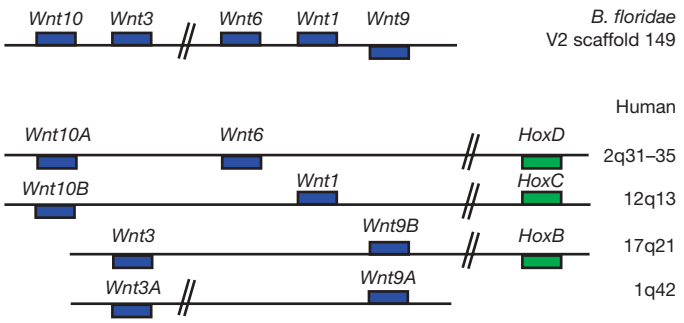


Figure 4 | Ancient developmental gene linkages. In the amphioxus genome, *Wnt6*, *Wnt1* and *Wnt9* form a compact gene cluster 2.5 Mb from *Wnt10* and *Wnt3* but all are on scaffold 149. Orthologues of four of these genes are also clustered in *Drosophila* (not shown), although *Wnt3* has been lost, as inferred from its presence in cnidarians. The human orthologues are on four chromosomes, and show disruption of gene clustering through duplication followed by gene loss. Linkages of Hox clusters to three of the human loci gives additional support for the large-scale duplication events involved. The three clusters that are linked to Hox clusters (as well as the four Hox clusters) fall in chromosome segments grouped in CLG 16, as does the Hox-bearing amphioxus scaffold. Genes drawn as boxes above the horizontal lines are transcribed from left to right; genes depicted below the lines are transcribed from right to left.

overlap with or are immediately adjacent to the 3′ or 5′ untranslated regions (UTRs) of human genes, and are probably conserved UTR elements. Four are adjacent to exons and probably represent conserved splicing enhancers. A single conserved noncoding element overlapped a highly conserved microRNA gene (*mir-10b* adjacent to the human *HOXD4* gene). The remaining 56 elements are of unknown function, but can be tested experimentally for enhancer activity[45].

Conclusions

The amphioxus sequence reveals key features of the genome of the last common ancestor of all chordates through comparison with the genomes of other animals. This ancestor probably lived before the Cambrian period, and gave rise to the chordate lineage that is represented today by modern cephalochordates such as amphioxus, as well as urochordates and vertebrates. Of the living lineages, the cephalochordates diverged first, before the split between the morphologically diverse urochordates and vertebrates. To a remarkable extent, the amphioxus genome appears to be a good surrogate for the ancestral chordate genome with respect to gene content, exon–intron gene structure and even chromosomal organization. The sequences of model ascidians with small genomes are by comparison simplified by gene loss, intron loss and genome rearrangement. Remarkably, modest levels of non-coding sequence have been conserved between amphioxus and human—the oldest conserved non-coding regions yet detected through direct sequence alignment—and may provide a tantalizing glimpse of the gene regulatory systems of the last common chordate ancestor.

Extensive conserved synteny between the genomes of amphioxus and various vertebrates lends unprecedented clarity and coherence to the history of genome-scale events in vertebrate evolution. The human and other jawed vertebrate genomes show widespread quadruple-conserved synteny relative to the amphioxus sequence, which extends earlier regional studies and provides a unified explanation for paralogous chromosomal regions in vertebrates. Our analysis thus provides conclusive evidence for two rounds of complete genome duplication on the jawed vertebrate stem. These genome duplications occurred after the divergence of urochordates but before the split between cartilaginous fish and bony vertebrates. The jawless vertebrates (for example, lamprey and hagfish) represent the only other chordate lineages that survive from this period, and at least the lamprey appears to have diverged between the two rounds of duplication, although the data still allow for an octoploid population as the progenitor of the jawless and jawed vertebrates. The detailed mechanism of these events—in particular, whether they occurred by allo- and/or auto-tetraploidizations, how closely spaced in time they were, and the precise nature of the rediploidization process—remain unknown. Although it is tempting to relate the genome duplications to specific morphological radiations in vertebrate evolution, the fossil record reflects a relatively steady diversification rather than a dramatic discontinuity of stem-group vertebrate forms[67].

The genomic features that are associated with organismal complexity, if such generic features exist at all, remain unknown[68]. It is tempting to speculate, however, that the creation of the ancestral jawed vertebrate genome by two rounds of genome duplication was a formative event in the early history of vertebrates that provided genomic flexibility through the duplication of coding and *cis*-regulatory sequences for the emergence of familiar developmental, morphological and physiological novelties such as chondrogenic and skeletogenic neural crest cells, the sclerotome (vertebral) compartment of the somites, elaborate hindbrain patterning, finely graded nervous system organization, and the elaborated endocrine system of vertebrates. Indeed, we find that genes involved in developmental signalling and gene regulation are significantly more likely to be retained in multiple copies in living species than genes overall, suggesting that diversified developmental regulation is correlated with the evolution of vertebrate novelties. This begs the question, dating back to Ohno, of how such duplicated genes became integrated into the biochemical and genetic networks of vertebrates. In a separate paper[45], we examine vertebrate biology in the light of the amphioxus genome data and the genome-scale duplication events on the vertebrate stem.

METHODS SUMMARY

Genome sequencing, assembly and annotation. High-quality sequence Sanger reads (7.3 million) were generated and assembled using JAZZ[69]. Protein-coding genes were annotated using EST, homology and *ab initio* methods as described previously[42,70].

Deuterostome relationships. Orthologous gene alignments were created using ClustalW[71] and Gblocks[72], and analysed with bayesian and maximum likelihood methods.

Intron evolution. The presence and absence of an intron at each of 5,337 orthologous coding positions was treated as a binary character in parsimony and bayesian analyses.

Construction of chordate linkage groups. Human chromosome segments and amphioxus scaffolds were clustered by orthologue distribution profile. The null hypothesis (orthologous genes randomly distributed across the two genomes) was evaluated using Fisher's exact test, with a Bonferroni correction for the total number of pairwise tests.

Decomposition of CLGs into independent products of duplication. The most parsimonious partitioning of the human chromosomal segments assigned to the CLG was obtained using a scoring system that included shared ohnologues and position in the multi-species synteny comparison.

NK quadruple conserved synteny. In addition to the genome-wide synteny analysis, a detailed manual curation was carried out on four v1 scaffolds (56, 124, 185 and 294) that make up the NK homeobox cluster in amphioxus. The 82 amphioxus genes correspond to 111 human genes enriched on chromosomes 4, 5, 8 and 10 (chi-squared test, $P \ll 0.001$), in agreement with the genome-wide analysis of CLG 7 (Supplementary Note 11).

Ancient developmental gene linkages. Orthology of homeodomain- and Wnt-containing genes was assigned from phylogenetic tree reconstruction using neighbour-joining and maximum likelihood approaches, supported by high bootstrap values.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 March; accepted 4 April 2008.

1. Costa, O. G. in *Cenni zoologici, ossia descrizione sommaria delle specie nuove di animali discoperti in diverse contrade del regno nell'anno 1834.* 49 (Azzolino, Napoli, 1834).
2. Yarrell, W. in *A History of British Fishes* 468–472 (Van Voorst, London, 1836).
3. Kowalevsky, A. Entwicklungsgeschichte des Amphioxus lanceolatus. *Mém. Acad. Imp. Sci. Saint-Pétersbourg* **11,** 1–17 (1866).
4. Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
5. Schmidtke, J., Weiler, C., Kunz, B. & Engel, W. Isozymes of a tunicate and a cephalochordate as a test of polyploidisation in chordate evolution. *Nature* **266,** 532–533 (1977).
6. Holland, P. W., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Development* (Suppl.) 125–133 (1994).
7. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
8. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
9. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.* **2,** 333–341 (2001).
10. Popovici, C., Leveugle, M., Birnbaum, D. & Coulier, F. Coparalogy: physical and functional clusterings in the human genome. *Biochem. Biophys. Res. Commun.* **288,** 362–370 (2001).
11. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of en bloc duplication in vertebrate genomes. *Nature Genet.* **31,** 100–105 (2002).
12. McLysaght, A., Hokamp, K. & Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31,** 200–204 (2002).
13. Lundin, L. G., Larhammar, D. & Hallbook, F. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J. Struct. Funct. Genomics* **3,** 53–63 (2003).
14. Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3,** e314 (2005).
15. Furlong, R. F. & Holland, P. W. Were vertebrates octoploid? *Phil. Trans. R. Soc. Lond. B* **357,** 531–544 (2002).
16. Stokes, M. D. & Holland, N. D. The lancelet. *Am. Sci.* **86,** 552–560 (1998).
17. Chen, J.-Y., Dzik, J., Edgecombe, G. D., Ramskold, L. & Zhou, G.-Q. A possible Early Cambrian chordate. *Nature* **377,** 720–722 (1995).
18. Chen, J.-Y. Huang, D.-Y. & Li, C.-W. An early Cambrian craniate-like chordate. *Nature* **402,** 518–522 (1999).

19. Conway Morris, S. & Whittington, H. B. The animals of the Burgess Shale. *Sci. Am.* **240,** 122–133 (1979).
20. Shu, D., Zhang, X. & Chen, L. Reinterpretation of *Yunnanozoon* as the earliest known hemichordate. *Nature* **380,** 428–430 (1996).
21. Mallatt, J. & Chen, J. Y. Fossil sister group of craniates: predicted and found. *J. Morphol.* **258,** 1–31 (2003).
22. Holland, L. Z. & Holland, N. D. Chordate origins of the vertebrate central nervous system. *Curr. Opin. Neurobiol.* **9,** 596–602 (1999).
23. Holland, N. D. & Chen, J. Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. *Bioessays* **23,** 142–151 (2001).
24. Yu, J. K. *et al.* Axial patterning in cephalochordates and the evolution of the organizer. *Nature* **445,** 613–617 (2007).
25. Garcia-Fernandez, J. & Holland, P. W. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* **370,** 563–566 (1994).
26. Castro, L. F. & Holland, P. W. Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. *Evol. Dev.* **5,** 459–465 (2003).
27. Luke, G. N. *et al.* Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc. Natl Acad. Sci. USA* **100,** 5292–5295 (2003).
28. Castro, L. F., Furlong, R. F. & Holland, P. W. An antecedent of the MHC-linked genomic region in amphioxus. *Immunogenetics* **55,** 782–784 (2004).
29. Panopoulou, G. & Poustka, A. J. Timing and mechanism of ancient vertebrate genome duplications–the adventure of a hypothesis. *Trends Genet.* **21,** 559–567 (2005).
30. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).
31. Sodergren, E. *et al.* The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314,** 941–952 (2006).
32. Howell, W. M. & Boschung, H. T. Jr. Chromosomes of the lancelet, *Branchiostoma floridae* (order Amphioxi). *Experientia* **27,** 1495–1496 (1971).
33. Small, K. S., Brudno, M., Hill, M. M. & Sidow, A. Extreme genomic variation in a natural population. *Proc. Natl Acad. Sci. USA* **104,** 5698–5703 (2007).
34. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
35. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99,** 803–808 (2002).
36. Britten, R. J., Rowen, L., Williams, J. & Cameron, R. A. Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl Acad. Sci. USA* **100,** 4661–4665 (2003).
37. Ivanova-Kazas, O. M. An essay on the phylogeny of lower chordates. *Trudy Sankt-Peterburgskogo Obshchestva Estestvoispytatelei* **84,** 1–158 (1995).
38. Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Mol. Biol. Evol.* **22,** 2275–2284 (2005).
39. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439,** 965–968 (2006).
40. Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444,** 85–88 (2006).
41. Cameron, C. B., Garey, J. R. & Swalla, B. J. Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. *Proc. Natl Acad. Sci. USA* **97,** 4469–4474 (2000).
42. Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317,** 86–94 (2007).
43. Edvardsen, R. B. *et al.* Hypervariable and highly divergent intron–exon organizations in the chordate *Oikopleura dioica*. *J. Mol. Evol.* **59,** 448–457 (2004).
44. Hughes, A. L. & Friedman, R. Loss of ancestral genes in the genomic evolution of *Ciona intestinalis*. *Evol. Dev.* **7,** 196–200 (2005).
45. Holland, L. Z. *et al.* The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* doi: 10.1101/gr.073676.107 (in the press).
46. Danchin, E. G. & Pontarotti, P. Towards the reconstruction of the bilaterian ancestral pre-MHC region. *Trends Genet.* **20,** 587–591 (2004).
47. Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* **17,** 1254–1265 (2007).
48. Venkatesh, B. *et al.* Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. *PLoS Biol.* **5,** e101 (2007).
49. Robinson-Rechavi, M., Boussau, B. & Laudet, V. Phylogenetic dating and characterization of gene duplications in vertebrates: the cartilaginous fish reference. *Mol. Biol. Evol.* **21,** 580–586 (2004).
50. University. of Washington. *The Sea Lamprey Genome Project* <http://genome.wustl.edu/genome.cgi?GENOME=Petromyzon%20marinus> (2007).
51. Escriva, H., Manzon, L., Youson, J. & Laudet, V. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol. Biol. Evol.* **19,** 1440–1450 (2002).
52. Kohn, M. *et al.* Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet.* **22,** 203–210 (2006).
53. Naruse, K. *et al.* A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* **14,** 820–828 (2004).
54. Woods, I. G. *et al.* The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* **15,** 1307–1314 (2005).
55. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).
56. Postlethwait, J. H. *et al.* Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* **10,** 1890–1902 (2000).
57. Panopoulou, G. *et al.* New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* **13,** 1056–1066 (2003).
58. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* **7,** R43 (2006).
59. Christoffels, A. *et al. Fugu* genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* **21,** 1146–1151 (2004).
60. Hoegg, S., Brinkmann, H., Taylor, J. S. & Meyer, A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* **59,** 190–203 (2004).
61. Brunet, F. G. *et al.* Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23,** 1808–1816 (2006).
62. Hellsten, U. *et al.* Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. *BMC Biol.* **5,** 31 (2007).
63. De Bodt, S., Maere, S. & Van de Peer, Y. Genome duplications and the origin of angiosperms. *Trends Ecol. Evol.* **20,** 591–597 (2005).
64. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
65. Pennacchio, L. A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).
66. Prochnik, S. E., Rokhsar, D. S. & Aboobaker, A. A. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev. Genes Evol.* **217,** 73–77 (2007).
67. Donoghue, P. C. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.* **20,** 312–319 (2005).
68. Valentine, J. W. Two genomic paths to the evolution of complexity in bodyplans. *Paleobiology* **26,** 513–519 (2000).
69. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu* rubripes. *Science* **297,** 1301–1310 (2002).
70. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
71. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
72. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).
73. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).
74. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
75. Lowe, C. J. *et al.* Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113,** 853–865 (2003).
76. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
77. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
78. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52,** 696–704 (2003).
79. Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods)* Version 4 (Sinauer Associates, Sinderland, Massachusetts, 2003).
80. Castro, L. F. & Holland, P. W. Fluorescent *in situ* hybridisation to amphioxus chromosomes. *Zoolog. Sci.* **19,** 1349–1353 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under contract number W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231, and Los Alamos National Laboratory under contract number DE-AC02-06NA25396. The Center for Integrative Genomics is supported by a grant from the Gordon and Betty Moore Foundation. D.S.R. acknowledges support from R. A. Melmon. This work was funded by grants from Ministerio de Educacin y Ciencia (J.G.-F.), MEXT, Japan (N.S., A.F., and Y.K.), the 21st Century and Global COEs at Kyoto University (N.S.), grant number P41LM from the National Library of Medicine (J.J. and V.V.K.), BBSRC (T.B. and D.E.K.F.) and the Wellcome Trust (P.W.H.H.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.S.R. (dsrokhsar@yahoo.com) or N.S. (satoh@ascidian.zool.kyoto-u.ac.jp).

METHODS

Genome sequence and assembly. High-quality sequence Sanger reads (7.3 million) were generated and assembled using JAZZ[69,73]. See Supplementary Note 2 for details of the genomic libraries, assembly methods and validation.

Annotation of protein-coding genes. Protein-coding genes were annotated by the JGI annotation pipeline as previously described[42,70]. See Supplementary Note 3 for a description of amphioxus-specific details. Gene models representing allelic pairs were identified using a combination of similarity of predicted peptide sequence and gene neighbourhood context (Supplementary Note 3).

Deuterostome relationships. Sets of orthologous genes were collected by grouping together mutual-best BLAST[74] hits between *N. vectensis* (sea anemone) and gene sets from other published deuterostome genomes plus expressed sequence data from the acorn worm *Saccoglossus kowalevskii*[75] and the sea lamprey *P. marinus* (Supplementary Note 5). Individual multiple alignments were created with CLUSTALW[71], were manually reviewed, trimmed with Gblocks[72], and concatenated. Orthologue sets (364) had representation from all genomes (alignment 1), and 1,090 had up to one missing (alignment 2). The alignments were analysed by bayesian and maximum likelihood methods, respectively, using MrBayes[76,77] and PHYML[78]. Supplementary Note 5 contains more details of the data sources, data compilation and analysis.

Intron evolution. A collection of 5,337 orthologous well-aligned coding sequence positions that contain an intron in at least one genome was analysed by weighted parsimony analysis with PAUP[79] and Bayesian analysis with MrBayes[76,77] (Supplementary Note 6).

Chordate gene families. Using predicted proteomes for human, chicken, stickleback, pufferfish, sea squirt, amphioxus, sea urchin, fruitfly and sea anemone (Supplementary Note 5), families ('clusters') of orthologous genes were constructed to represent the ancestral gene complements of the tetrapod, teleost, jawed vertebrate, 'olfactores' (that is, vertebrates plus urochordates), chordate and deuterostome ancestors, as described previously[42], with modifications described in Supplementary Note 7.

Chromosome segmentation. The human, chicken and stickleback chromosomes were segmented iteratively by comparison to one another and to the scaffolds of the *Fugu* genome assembly. See Supplementary Note 8 for complete details.

Construction of chordate linkage groups. For whole-genome synteny analysis, orthology between genomes was based on *c*-score (BLAST score/best BLAST score) clustering as previously described[42], with a *c*-score threshold of 0.75 when comparing human and amphioxus, and 0.95 when comparing human to other vertebrates. To define initial CLGs (Supplementary Fig. 64), human chromosome segments and amphioxus scaffolds were clustered using the same method as for chromosome segmentation, with a correlation threshold of 0.25. Statistical significance of orthologue concentration between regions of one genome and another was computed with Fisher's exact test with a Bonferroni correction for the total number of pairwise tests (see Supplementary Note 8 for additional details).

Fluorescent *in situ* hybridization. Chromosome preparation was performed as described previously[80] with modifications described in Supplementary Note 8.

Multi-species synteny comparison. The clustering protocol described above for human–amphioxus was repeated for human–*Fugu*, human–stickleback and human–*Nematostella* to define clusters of scaffolds or chromosome segments (a 'cluster set') for each genome based on comparison to human. All pairs of human chromosome segments were compared to each cluster set, and for each set were classified as having conserved synteny to the same cluster (coded with '1'), having conserved synteny (only) to different clusters (coded with '0'), or having indeterminate conserved synteny if one or both human segments lack significant conserved synteny to any cluster in the cluster set (coded with '?'). The complete results are represented as a colour-coded matrix in Supplementary Fig. 9.

Identification of ohnologues. Operationally, we define a pair of human genes as ohnologues (paralogues descendent from 2*R* in vertebrate evolution) if they are found in the same chordate gene family (excluding large gene families with more than ten members) and are ancient paralogues differing by more than 0.2 transversions per site at synonymous positions. (We use transversions rather than the more common total substitutions because transversions occur more slowly and therefore show less saturation at the timescales of interest.)

Decomposition of CLGs into independent products of duplication. For each CLG, all partitionings of the human chromosomal segments assigned to the CLG were tabulated. Each partitioning was assigned a score as follows: $+1$ for each pair of segments from different partitions with a significant number of predicted ohnologues between them; -1 for each pair of segments from different partitions without a significant number of ohnologues between them. Among the partitionings with the maximum score, ties were broken by using the multi-species synteny comparison results: a score of $+\varepsilon$ for each pair of segments from different partitions coloured red or orange, and a score of $-\varepsilon$ for each pair of segments from different partitions where multi-synteny comparison indicates the two segments were one segment in the jawed vertebrate ancestor (coloured blue or purple in Supplementary Figs 9 and 10), where epsilon is a positive number much less than 1.

Timing of genome duplications. Genes from pufferfish, lamprey and sea squirt ('*X*') were aligned to pairs of human ohnologues ('Hs1' and 'Hs2') and the orthologous amphioxus gene; phylogenetic position was considered resolved if it has at least 50% maximum likelihood bootstrap support (Supplementary Note 9 and Supplementary Fig. 11).

Ancient developmental gene linkages. Genes were identified by tBLASTn against version 1.0 of the *B. floridae* genome assembly with vertebrate and invertebrate homeodomain and Wnt sequences. Orthology was assigned from phylogenetic tree reconstruction using neighbour-joining and maximum likelihood approaches. Support for nodes was assessed by bootstrapping; all gene families were recovered with high support. Human data from Ensembl release 47 were used. Supplementary Table 3 lists the genes and gene models examined.

Functional categorization of retained duplicate genes. PANTHER functional annotations were mapped to inferred ancestral chordate genes, and subsets of these genes were analysed for enrichment in functional categories by methods previously described for the analysis of ancestral eumetazoan genes[42]. Because functional annotations overlap, the category of 'developmental processes' is itself dominated by genes associated with signal transduction and transcriptional regulation.

Conserved non-coding element and expression analysis. For conserved non-coding element and expression analysis, see Supplementary Note 10.

ARTICLES

PML targeting eradicates quiescent leukaemia-initiating cells

Keisuke Ito[1,3,4], Rosa Bernardi[1,3,4], Alessandro Morotti[1,3,4], Sahoko Matsuoka[5], Giuseppe Saglio[6], Yasuo Ikeda[5], Jacalyn Rosenblatt[2], David E. Avigan[2], Julie Teruya-Feldstein[4] & Pier Paolo Pandolfi[1,3,4]

The existence of a small population of 'cancer-initiating cells' responsible for tumour maintenance has been firmly demonstrated in leukaemia. This concept is currently being tested in solid tumours. Leukaemia-initiating cells, particularly those that are in a quiescent state, are thought to be resistant to chemotherapy and targeted therapies, resulting in disease relapse. Chronic myeloid leukaemia is a paradigmatic haematopoietic stem cell disease in which the leukaemia-initiating-cell pool is not eradicated by current therapy, leading to disease relapse on drug discontinuation. Here we define the critical role of the promyelocytic leukaemia protein (PML) tumour suppressor in haematopoietic stem cell maintenance, and present a new therapeutic approach for targeting quiescent leukaemia-initiating cells and possibly cancer-initiating cells by pharmacological inhibition of PML.

The existence of cancer-initiating cells (CICs), a minor subpopulation of cells responsible for tumour initiation and maintenance, was proposed over 40 years ago[1]. In leukaemia in particular, increasing evidence suggests that out of the bulk of leukaemic cells only a rare population of leukaemia-initiating cells (LICs) propagates the disease[2–7]. LICs are rare and share many properties of normal haematopoietic stem cells (HSCs), such as self-renewal, pluripotency and quiescence[4–6]. A fundamental problem in treating leukaemia lies in the fact that LICs remain untouched by both conventional chemotherapy and even by targeted therapies[7]. The quiescent LIC subpopulation is thought to be particularly resistant to drugs that would normally target cells in active DNA replication[7]. Hence leukaemia relapse may occur because therapies eliminate proliferating cells that constitute the bulk of the tumour, but fail to eradicate quiescent LICs that can reinitiate malignancy after a period of latency. Therefore, development of new therapeutic approaches targeting CICs and LICs may have a profound impact on cancer eradication.

Chronic myeloid leukaemia (CML) is one of the most extensively investigated and paradigmatic stem cell disorders[7]. It is characterized by the presence of the Philadelphia chromosome, which results from a chromosomal translocation between the *BCR* gene on chromosome 22 and the *ABL* gene on chromosome 9 (refs 8, 9). This translocation generates the fusion protein BCR–ABL which displays constitutive kinase activity[10]. The tyrosine kinase inhibitor imatinib markedly improves the prognosis of CML patients[11,12]; however, imatinib preferentially targets dividing cells, whereas non-dividing leukaemic cells are resistant to imatinib-mediated apoptosis[6]. Surviving leukaemia stem and progenitor cells are a potential source for relapse. This is demonstrated by the fact that, if therapy is discontinued, the disease inevitably relapses in most cases, including those showing good responses without signs of disease progression[13–18].

The *PML* gene, which is involved in the t(15;17) chromosomal translocation of acute promyelocytic leukaemia (APL), encodes a protein localizing to PML nuclear bodies, a subnuclear macromolecular structure[19]. PML functions as a tumour suppressor that controls fundamental processes such as apoptosis, cellular proliferation and senescence[20,21]. Recent data demonstrated that PML is involved in neoangiogenesis and acts as a negative regulator of mTOR[22]. However, its role in stem cell biology has not been investigated. Here, we studied the role of PML in HSC and LIC biology and obtained unexpected data that have implications for the eradication of LICs and CICs in human cancer.

Loss of PML predicts favourable outcome in CML

To understand whether PML expression is modulated during haematopoiesis, we analysed Pml protein levels in various haematopoietic cell lineages in the mouse. To detect Pml levels in rare HSCs, we sorted different cell lineages directly into a sample buffer. Western blot analysis showed that Pml is highly expressed in the HSC compartment (Fig. 1a). Immunofluorescence analysis also showed increased numbers of PML nuclear bodies in HSCs compared to committed cells (Supplementary Fig. 1a). *Pml* mRNA levels were also higher in the HSC population, indicating that Pml expression during haematopoiesis is regulated at the transcriptional level (Fig. 1b). High PML expression in the HSC compartment was also observed in primary human bone marrow samples (Supplementary Fig. 1b, c).

We next evaluated PML expression in samples of patients with haematopoietic malignancies. Loss of PML is frequently observed in human cancers such as prostate and lung cancer[23,24]. However, most CML chronic phase samples expressed high levels of PML (Fig. 1c). Moreover, PML expression was barely detected in differentiated neutrophils whereas abundant PML expression was seen in blasts expressing CD34 (Fig. 1c and Supplementary Fig. 1d). An unexpected association was found between PML positivity and clinical outcome: CML patients with low PML expression displayed higher complete molecular response (CMR) and complete cytogenetic response (CCyR) compared with patients with high PML expression (Fig. 1d, e). Furthermore, low PML expression was strikingly predictive of better overall survival in CML (Supplementary Fig. 2). These results indicate that in CML, low PML expression

[1]Cancer Genetics Program, Beth Israel Deaconess Cancer Center, Departments of Medicine and Pathology, [2]Division of Hematology and Oncology, Beth Israel Deaconess Medical Center, Harvard Medical School, New Research Building, 330 Brookline Avenue, Boston, Massachusetts 02215, USA. [3]Cancer Biology and Genetics Program, [4]Department of Pathology, Memorial Sloan-Kettering Cancer Center, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA. [5]Division of Hematology, Department of Internal Medicine, Keio University School of Medicine, 35 Shinano-machi, Shinjuku-ku, Tokyo 160-8582, Japan. [6]Division of Hematology and Internal Medicine, Department of Clinical and Biological Sciences, University of Turin, Turin 10043, Italy.

predicts a better clinical outcome contrary to what was observed in prostate cancer and other solid tumours[23,24]. These results prompted us to analyse the role of PML in haematopoiesis.

Pml is required for HSC maintenance

Genetic loss of *Pml* did not induce significant changes in the number of haematopoietic cells in peripheral blood (data not shown) and in the quantity or quality of progenitors in the bone marrow in 8-week-old mice (Supplementary Fig. 3a–d). However, an increased number of cells in the c-Kit$^+$Sca-1$^+$Lin$^-$ (KSL) stem cell compartment was found in 8-week-old *Pml*$^{-/-}$ mice (Fig. 2a). In particular, the number of long-term repopulating HSCs measured as CD34$^-$ and Thy1^{low} KSL cells was significantly higher in 8-week-old *Pml*-deficient mice (Fig. 2a). The proportion of cells in G0 among KSL and CD34$^-$ KSL cells, as evaluated by pyronin Y staining[25], was markedly lower in *Pml*$^{-/-}$ mice than in wild-type mice (Fig. 2b), indicating that *Pml*$^{-/-}$ HSCs are not quiescent. Consistent with these data, the number of colony-forming cells from *Pml*$^{-/-}$ KSL cells was higher than wild-type KSL cells after short-term culture on stromal cells (less than 2 weeks). However, the number of colonies from *Pml*$^{-/-}$ KSL cells decreased significantly after 6 weeks of culture (Fig. 2c). These results suggest that increased cycling of *Pml*$^{-/-}$ HSCs results in their exhaustion. To assess the repopulating ability of *Pml*$^{-/-}$ HSCs *in vivo*, we performed a competitive reconstitution assay. Flow cytometric analysis revealed that *Pml*$^{-/-}$ KSL cells contributed to haematopoietic reconstitution more than competitor cells 4 weeks after transplantation (Fig. 2d); however, the percentage of *Pml*$^{-/-}$ KSL cells significantly decreased 16 weeks after transplantation (Fig. 2d). These results indicate that Pml acts to maintain HSCs and that *Pml*$^{-/-}$ HSCs lack long-term repopulating capacity. This defect affected both myeloid as well as B and T lineages (Supplementary Fig. 4a). Cell cycle analysis of recipient bone marrow revealed that more HSCs from *Pml*$^{-/-}$ donors were cycling than those from wild-type donors (Fig. 2e), indicating that *Pml*$^{-/-}$ HSCs are not quiescent in the bone marrow of recipient mice. Analysis of chimaerism revealed that all haematopoietic lineages from *Pml*$^{-/-}$ donors were affected, but the greatest reductions were seen in the HSC compartment (Fig. 2f). In addition, the contribution of *Pml*$^{-/-}$ HSCs to more committed cells was more significantly impaired at later time points after transplantation (Supplementary Fig. 4b).

The impact of *Pml* deficiency on long-term repopulation was determined by carrying out a second competitive bone marrow transplantation (BMT). *Pml*$^{-/-}$ donor-derived cells could not reconstitute the bone marrow of recipient mice in the second BMT (Supplementary Fig. 4c). Consistent with these data, defects in progenitor function were observed in *Pml*$^{-/-}$ donor-derived cells after BMT (Supplementary Fig. 4d).

To assess HSC function under normal homeostatic conditions, we examined the effect of *Pml* deficiency on haematopoiesis in older mice. Older *Pml*$^{-/-}$ mice exhibited a progressive decrease in cellularity, with a mean ratio of bone marrow mononuclear cells compared to wild-type mice of 0.67 ± 0.07 and 0.50 ± 0.06 at 12 and 18 months. Additionally, in contrast to the increased number of KSL cells seen at 2 months of age, a significant reduction of HSCs was evident in *Pml*$^{-/-}$ bone marrow at 18 months (Fig. 2g), accompanied by marked progenitor dysfunction (Supplementary Fig. 4e). Finally, repopulating cells from 18-month-old *Pml*$^{-/-}$ bone marrow were not detected in recipient mice even 4 weeks after transplantation (Supplementary Fig. 4f). Thus, our data indicate that chronic *Pml* deficiency *in vivo* results in progressive impairment of HSC function due to defective maintenance of quiescence.

PML is indispensable for LIC maintenance

LICs have notable mechanistic similarities to normal stem cells[2,4,26]. Therefore, because high PML expression was seen in CML blasts (Fig. 1c), we investigated the function of PML in CML LICs.

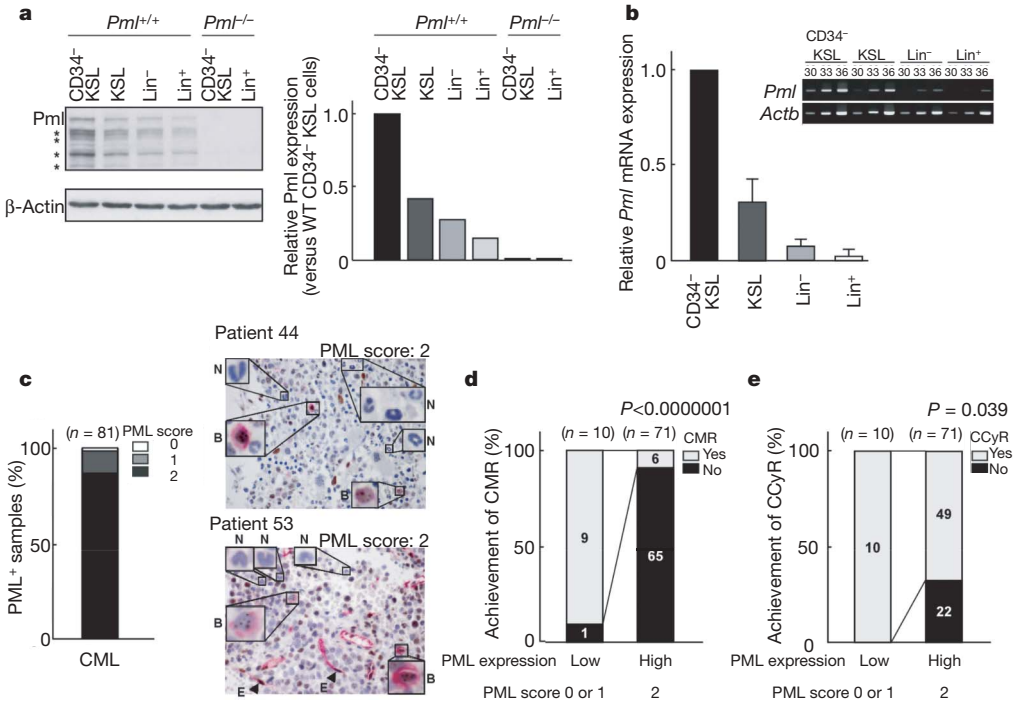


Figure 1 | PML is highly expressed in HSCs and CML. a, Fractionated mouse haematopoietic cells were flow-sorted into protein sample buffer and immunoblotted with anti-Pml antibody. Representative blots are shown in the left panel and relative Pml protein level normalized to β-actin are shown in the right panel. Asterisks indicate Pml isoforms. **b**, Levels of *Pml* and *Actb* transcripts were measured by q-RT–PCR in haematopoietic cells. The bar graph represents normalized expression of *Pml* mRNA. Experiments were performed twice and a representative result is presented. Error bar indicates s.d. High Pml expression was also confirmed by PCR (inset). Aliquots were taken at the end of the indicated cycles. **c**, Bone marrow samples of CML patients in chronic phase ($n = 81$) were stained with anti-PML antibody (brown) and anti-CD34 (red). The left graph shows the percentage of PML-positive samples and representative cases are on the right (arrowheads indicate endothelial cells (E) as a positive control). Insets show PML-staining in blasts (B) and differentiated neutrophils (N). **d**, **e**, Higher CMR (**d**) and CCyR (**e**) were observed in chronic-phase CML patients with low PML expression. *P*-values were generated by a chi-squared test. Absolute numbers are also indicated.

Pml$^{+/+}$ and *Pml*$^{-/-}$ bone marrow cells were transduced with *p210*$^{BCR\text{-}ABL}$, and then cultured on stromal cells to enrich LICs. Pyronin Y staining of KSL cells revealed a significant reduction in the number of quiescent cells in *Pml*-deficient cells compared to wild-type cells (Fig. 3a). Consistent with these data, *Pml* null LICs showed increased colony-forming capacity after short-term culture but remarkable reduction in colony number after long-term culture on stromal cells when compared with wild-type LICs (Fig. 3b). To investigate the function of Pml in LICs *in vivo*, we serially transplanted bone marrow cells transduced with *p210*$^{BCR\text{-}ABL}$ to recipient mice every 2 weeks. Retroviral transduction of *p210*$^{BCR\text{-}ABL}$ results in transformation of bone marrow cells, resulting in CML-like disease[27]. In the first BMT, *Pml*$^{-/-}$ LICs promoted earlier CML-like disease in recipient mice than wild-type LICs (Supplementary Fig. 5a, b). When cell cycle status was investigated, significantly fewer *Pml*-deficient than wild-type LICs appeared in G0 in recipient mice (Fig. 3c).

In the second BMT, no significant difference in survival was observed (Supplementary Fig. 5c). In the third serial BMT, however, *Pml*$^{-/-}$ LICs failed to generate CML-like disease, contrary to wild-type LICs (Fig. 3d–f and Supplementary Fig. 5d). In addition, minimal residual disease was not detected in recipient mice transplanted with *Pml*$^{-/-}$ LICs (Fig. 3g). Remarkably, wild-type LICs retained the potential to develop CML-like disease even in the fourth serial BMT (Supplementary Fig. 5e). These results indicate that *Pml*-deficient LICs undergo intensive cell cycling, resulting in impairment of LIC maintenance.

As_2O_3 reversibly decreases PML expression in HSCs

The inorganic arsenite arsenic trioxide (As_2O_3) has been used as a therapeutic agent for centuries[28]. From the 1700s through to the early 1900s, arsenicals were a mainstay in the treatment of leukaemia[29]. The dramatic ability of arsenic to cure APL was reported in the mid-1990s[30–32]. Arsenic has been shown to target PML for degradation[33]. Indeed, when we analysed the effect of As_2O_3 treatment on mouse HSCs, we found that it reversibly decreased Pml expression in the HSC compartment *in vitro* (Fig. 4a). Reduction of Pml expression by As_2O_3 markedly attenuated colony-formation ability of wild-type KSL cells compared to control cells after 6 weeks on stromal cells, whereas an increase in colony formation was observed after short-term culture (Fig. 4b). Notably, As_2O_3 treatment did not affect colony formation from *Pml*$^{-/-}$ HSCs (Fig. 4b), indicating that this effect is mostly Pml-dependent. *In vivo* treatment with As_2O_3 also reversibly reduced Pml expression (Supplementary Fig. 6), and resulted in impaired HSC quiescence and an increase in the number of KSL cells (Fig. 4c, d). Furthermore, the number of long-term repopulating HSCs was significantly more elevated after As_2O_3 treatment (Fig. 4e).

Rapamycin rescues the phenotype of *Pml*$^{-/-}$ HSCs and LICs

Recent data have demonstrated that PML acts as a repressor of neoangiogenesis by repressing mTOR activity in conditions of hypoxia[22]. Because mTOR has an essential role in HSC maintenance

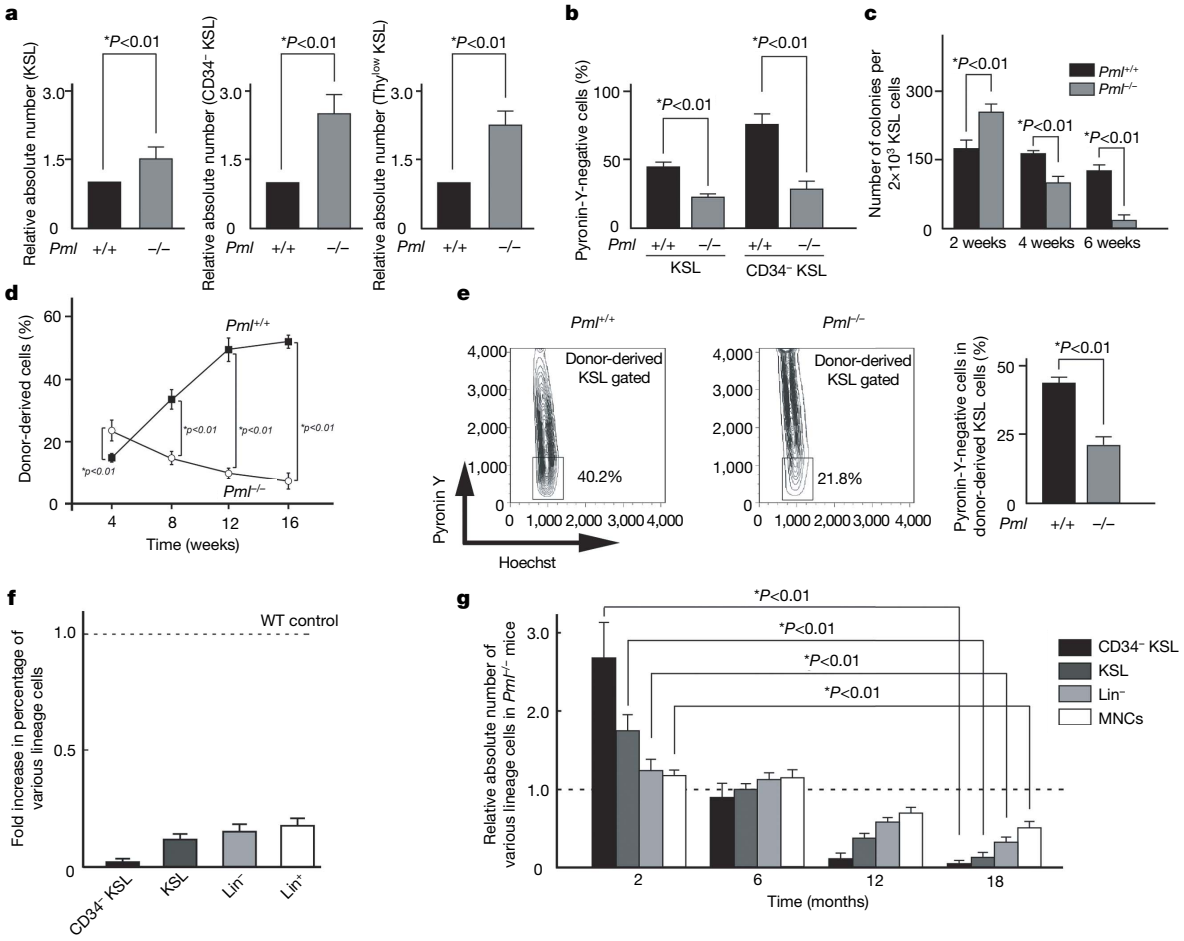


Figure 2 | Pml is essential for HSC maintenance. **a**, Relative numbers of KSL (left), CD34$^-$ KSL (middle) and Thy1^{low} KSL (right) cells in *Pml*$^{-/-}$ and *Pml*$^{+/+}$ bone marrow at 8 weeks ($n = 3$). **b**, Pyronin-Y-negative cells in KSL cells and CD34$^-$ KSL cells of *Pml*$^{+/+}$ or *Pml*$^{-/-}$ mice ($n = 3$). **c**, Colony-forming ability of wild-type and *Pml*$^{-/-}$ KSL cells after long-term culture ($n = 3$). **d**, Reconstitution of wild-type and *Pml*$^{-/-}$ bone marrow cells after competitive transplantation assay. **e**, Frequency of wild-type and *Pml*$^{-/-}$ quiescent cells in recipient mice. Right: representative flow cytometry data. Left: mean percentages of pyronin-Y-negative cells in donor-derived KSL population. **f**, Relative percentage of donor-derived cells in the bone marrow of recipient mice 4 months after transplantation ($n = 3$). **g**, Relative numbers of fractionated haematopoietic cells in *Pml*$^{-/-}$ mice at the indicated ages normalized over wild-type mice ($n = 3$). MNCs, mononuclear cells. All error bars indicate s.d.

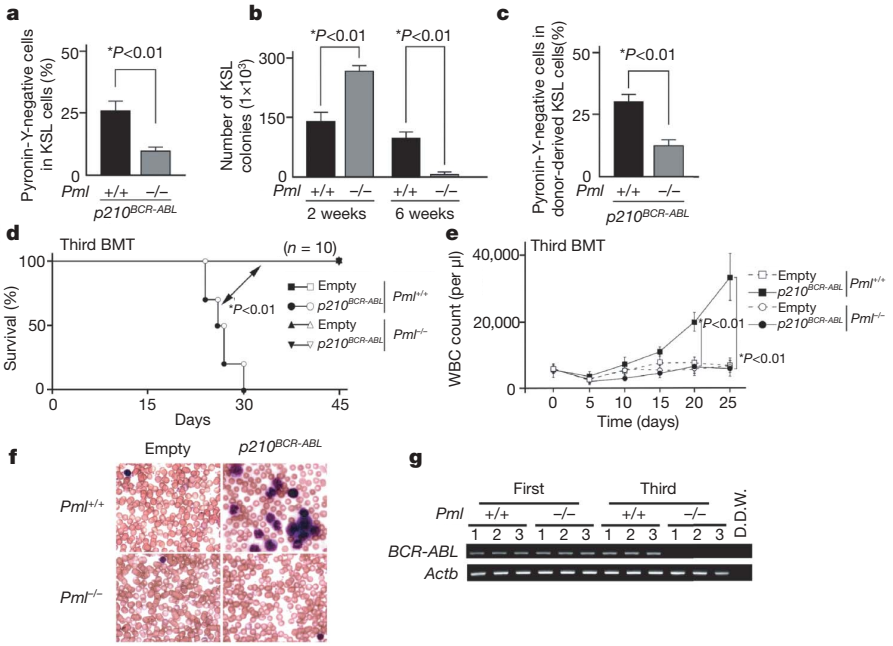


Figure 3 | Pml is essential for LIC maintenance. **a**, $Pml^{+/+}$ or $Pml^{-/-}$ bone marrow cells transduced with $p210^{BCR\text{-}ABL}$ were co-cultured with stromal cells for 2 weeks. Data shown are mean percentages of pyronin-Y-negative cells in KSL cells. **b**, Colony formation after long-term culture of $p210^{BCR\text{-}ABL}$-transduced bone marrow cells ($n = 3$). **c**, Cell cycle status of donor-derived KSL cells transduced with $p210^{BCR\text{-}ABL}$ 2 weeks after BMT. **d**, **e**, Survival of recipient mice receiving transduced bone marrow cells from $Pml^{+/+}$ or $Pml^{-/-}$ mice in third round BMT (**d**). Log-rank statistical analysis was performed to obtain *P* values. White blood cell (WBC) counts at indicated times after BMT are shown (**e**). **f**, Smears of peripheral blood (PB) in third round BMT stained with Wright–Giemsa. **g**, Minimal residual disease in third BMT recipient mice with wild-type or $Pml^{-/-}$ bone marrow cells overexpressing $p210^{BCR\text{-}ABL}$ was analysed by nested PCR in three randomly selected recipient mice. All error bars indicate s.d.

as well as leukaemogenesis[34,35], we examined mTOR activity in $Pml^{-/-}$ HSCs. Increased activity of mTOR was observed in *Pml*-deficient compared to wild-type HSCs (Supplementary Fig. 7a). *In vitro* treatment with the mTOR inhibitor rapamycin substantially restored colony-forming capacity in long-term cultures of $Pml^{-/-}$ HSCs, whereas it did not affect wild-type HSCs (Supplementary Fig. 7b). *In vivo* administration of rapamycin increased the quiescence of $Pml^{-/-}$ HSCs (Supplementary Fig. 7c) and resulted in decreased numbers of $Pml^{-/-}$ long-term HSCs (Supplementary Fig. 7d). Moreover, rapamycin treatment dramatically restored the capacity

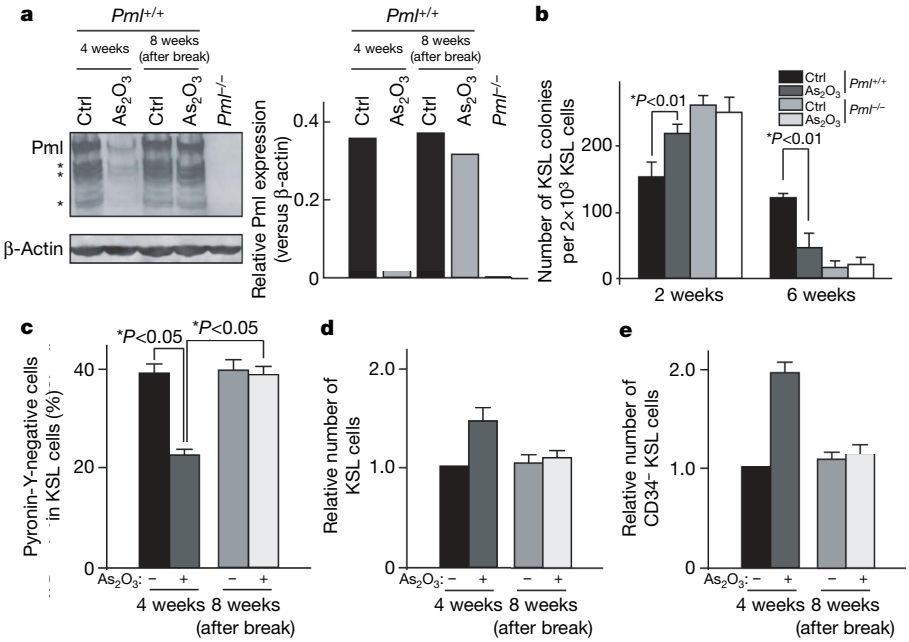


Figure 4 | Reduction of Pml by As_2O_3 treatment abrogates maintenance of HSC quiescence. **a**, KSL cells from 8-week-old mice were sorted and co-cultured with stromal cells and As_2O_3 for 4 weeks. As_2O_3 treatment was discontinued and co-culture continued for 4 weeks without treatment (8 weeks (after break)). Proteins from sorted KSL cells were analysed by western blot (left). Normalized Pml protein levels versus β-actin is shown on the right. **b**, $Pml^{+/+}$ and $Pml^{-/-}$ KSL cells were cultured on stromal cells with As_2O_3 for the indicated weeks and tested for colony formation ($n = 3$). **c**–**e**, As_2O_3 reversibly inhibits quiescence of normal HSCs *in vivo*. Mice were treated with As_2O_3 from 8 to 12 weeks (4 weeks) and left untreated from week 12 to 16 (8 weeks (after break)). Cell cycle status of HSCs was analysed by pyronin Y staining (**c**). Mean numbers of KSL (**d**) and $CD34^-$ KSL cells (**e**) are also shown. All error bars indicate s.d.

of *Pml*-deficient HSCs to provide long-term bone marrow reconstitution to irradiated mice (Supplementary Fig. 7e, f). These results indicate that Pml has an important role in the maintenance of HSCs by repressing mTOR activity.

We next examined the effect of rapamycin on *Pml*-deficient LICs. Administration of rapamycin significantly prevented the exhaustion of $Pml^{-/-}$ LICs, leading to restored colony-forming capacity in long-term culture (Supplementary Fig. 7g). In addition, $Pml^{-/-}$ LICs treated with rapamycin gave rise to CML-like disease in the third serial BMT, although in the first BMT disease onset was delayed (Supplementary Fig. 7 h–j). Notably, rapamycin also accelerated CML-like disease by wild-type LICs in the third serial BMT (Supplementary Fig. 7i). These results suggest that Pml acts as a repressor of mTOR activity in LICs and mTOR super-activation impairs LIC-maintenance.

PML downregulation is effective for LIC eradication

Interventions that enhance cycling of quiescent, chemotherapy-insensitive LICs are expected to facilitate their elimination. Therefore we investigated the therapeutic effect of As_2O_3-mediated PML reduction in LICs. As_2O_3 treatment significantly decreased the number of quiescent LICs without inducing apoptosis (Fig. 5a and Supplementary Fig. 8). Consistently, long-term culture-initiating cell assays revealed a remarkable inhibitory effect of As_2O_3 on LIC maintenance (Fig. 5b).

To verify whether As_2O_3-induced cycling could increase the pro-apoptotic effect of chemotherapy on LICs, we combined arsenic and cytosine arabinoside (Ara-C) treatment. Arsenic followed by Ara-C exposure significantly increased the efficacy of Ara-C-mediated induction of apoptosis, resulting in eradication of LICs even 4 weeks after treatment discontinuation (Fig. 5c and Supplementary Fig. 9a).

To analyse the effect of combination therapy on the persistence of long-term repopulating LICs, we treated LICs *ex vivo*, and next carried out serial transplantation assays. In the second round of BMT, mice transplanted with LICs treated with Ara-C succumbed around 20 days after BMT (Supplementary Fig. 9b). However, when donor LICs were treated with As_2O_3 and Ara-C, CML-like disease was not observed in recipient mice up to 40 days after BMT. These results

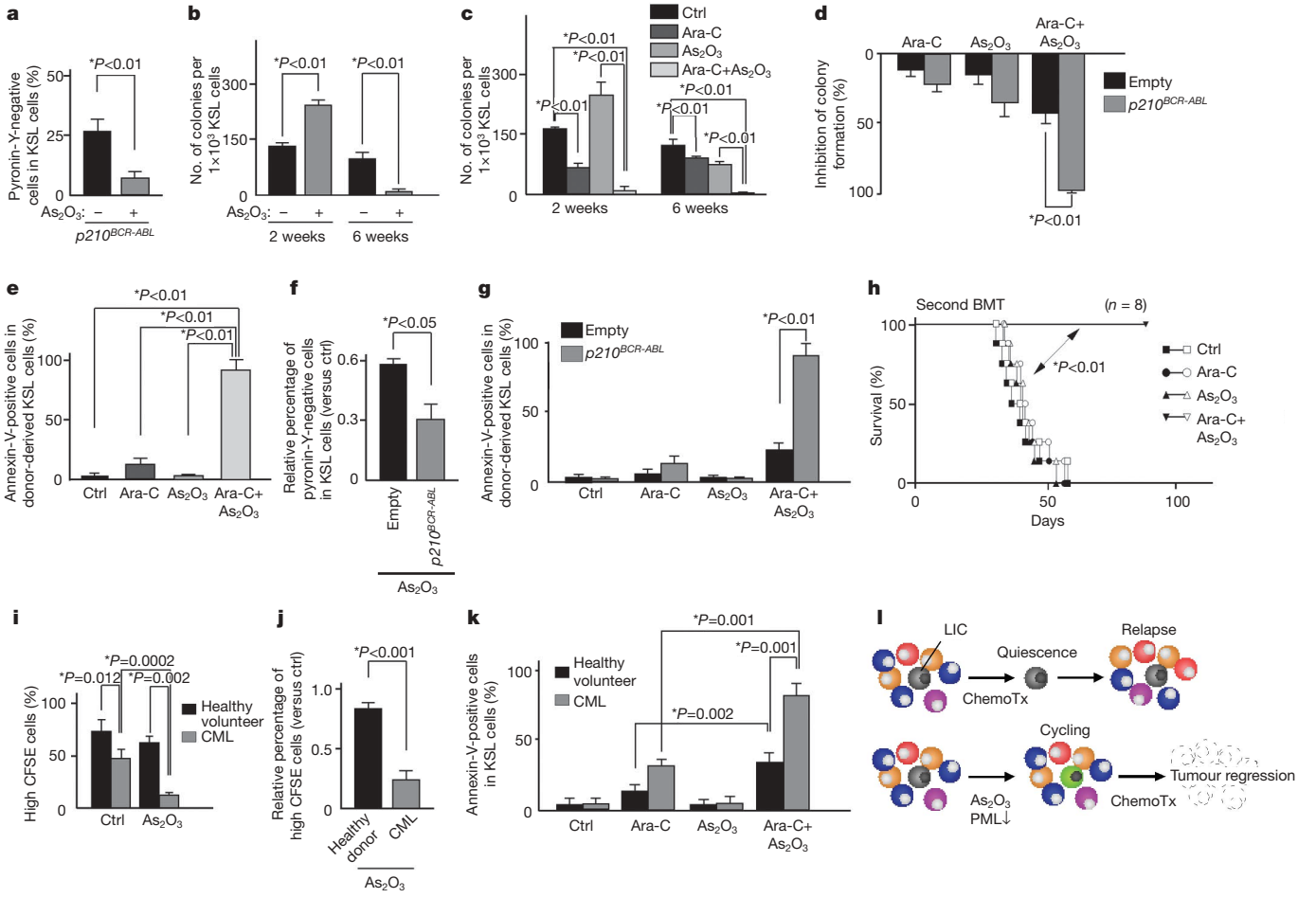


Figure 5 | Combination therapy with Ara-C and As_2O_3 eliminates LICs. **a**, Cell cycle analysis of bone marrow cells transduced with $p210^{BCR\text{-}ABL}$ co-cultured with stromal cells and As_2O_3 for 2 weeks. **b**, Colony formation of transduced KSL cells after long-term culture with As_2O_3 ($n = 3$). **c**, Transduced KSL cells co-cultured with stromal cells were treated with As_2O_3 for 9 days and with As_2O_3 and Ara-C for 5 days (2 weeks). Treatment was discontinued and co-culture continued for 4 weeks (6 weeks). Results are mean colony numbers ($n = 3$). **d**, Colony formation of As_2O_3- and Ara-C-treated bone marrow cells compared to untreated bone marrow cells infected by empty vector or $p210^{BCR\text{-}ABL}$ *in vitro* ($n = 3$). **e**, Apoptosis in donor-derived transduced KSL cells in recipient mice treated with As_2O_3 and Ara-C ($n = 3$). **f**, Quiescence of KSL cells transduced with empty vector or $p210^{BCR\text{-}ABL}$ after As_2O_3 treatment *in vivo*. **g**, Annexin-V staining of HSCs or LICs in recipient mice treated with As_2O_3 and Ara-C ($n = 3$). **h**, Survival of recipient mice transplanted with transduced bone marrow cells at the second round of BMT. The *P* value was obtained by log-rank statistical analysis. **i**, Percentage of non-dividing cells in $Lin^-CD34^+CD38^-$ cells from patients with CML and healthy volunteers cultured for 3 days ($n = 3$). **j**, Relative percentages of non-dividing cells in cells treated with As_2O_3 versus untreated cells. **k**, Apoptosis in $Lin^-CD34^+CD38^-$ cells co-cultured with As_2O_3 and Ara-C ($n = 4$). **l**, A model for As_2O_3-induced sensitization to therapy. Conventional chemotherapy (ChemoTx) does not affect quiescent LICs. As_2O_3 abrogates LIC maintenance by reducing PML levels. Combining an anti-leukaemic treatment with As_2O_3 increases the sensitivity of LICs to chemotherapy and results in tumour regression. All error bars indicate s.d.

indicate that induction of cell cycle entry by As_2O_3 remarkably enhances the effect of Ara-C, leading to a significant increase in survival. Furthermore, we not only observed a marked survival advantage, but also a complete cure in more than half of recipient mice (Supplementary Fig. 9b). Notably, residual disease was not detected in these mice (Supplementary Table 1).

Notably, we observed that there are fewer quiescent cells in $p210^{BCR\text{-}ABL}$ expressing cells than in control KSL cells (Supplementary Fig. 10a, b), indicating that the reservoir of quiescent cells is higher in HSCs than LICs. Moreover, exit from quiescence induced by As_2O_3 treatment was significantly more profound in LICs than in HSCs (Supplementary Fig. 10c). Similarly, a more profound reduction in LICs quiescence was observed in the *Pml* null setting (Supplementary Fig. 10d). Taken together, these findings suggest that LICs are more sensitive to induction of cell cycle entry by As_2O_3 than HSCs. Consequently, through induction of apoptosis, combination therapy with As_2O_3 and Ara-C affected LIC function significantly more than normal HSC function in long-term culture assays (Fig. 5d and Supplementary Fig. 10e).

We next investigated the effect of combination therapy on LIC maintenance in a serial transplantation model. After BMT of cells transduced with $p210^{BCR\text{-}ABL}$ or empty vector, *in vivo* administration of As_2O_3 to recipient mice followed by Ara-C treatment induced remarkable apoptosis in LICs (Fig. 5e). Transduced KSL cells were also much more prone to cell cycle entry by As_2O_3 treatment than control KSL cells in the transplantation model (Fig. 5f) and significantly more apoptosis was observed in LICs treated with As_2O_3 and Ara-C than in HSCs (Fig. 5g).

Consequently, complete cure with no detectable residual disease was achieved in all recipient mice treated with combination therapy in the second and third round of BMT (Fig. 5h, Supplementary Fig. 11a, b and Supplementary Table 2). In the third BMT As_2O_3 alone caused longer survival than Ara-C alone (Supplementary Fig. 11b), indicating that inhibition of maintenance may be more effective for tumour regression than targeting cycling cells with chemotherapy.

Finally, we analysed the impact of As_2O_3 treatment on stem cells isolated from human CML patients. First, similarly to murine LICs, fewer quiescent cells were observed in LICs from CML patients compared to HSCs from healthy volunteers (Fig. 5i). In addition, As_2O_3 treatment induced cell cycle induction more remarkably in LICs than HSCs, and was accompanied by downregulation of PML (Fig. 5i, j and Supplementary Fig. 12a, b). A more pronounced exit from quiescence in LICs compared to HSCs was also confirmed at the single cell level (Supplementary Fig. 12c). Finally, *in vitro* pre-treatment of human CML LICs with As_2O_3 followed by Ara-C induced significantly more apoptosis than Ara-C treatment alone, and a more profound apoptotic response was observed in LICs than in HSCs from healthy volunteers (Fig. 5k).

Discussion

It has been suggested that a rare population of leukaemic cells with stem characteristics (LICs) sustains the development of at least some form of leukaemia, including CML[2]. These cells are unresponsive to therapy and have been suggested as a cause of disease relapse[6,7,36]. Therefore, therapeutic strategies that target LICs are necessary to eradicate residual disease and to prevent leukaemia relapse.

We used a CML mouse model to analyse LIC function in the absence of the tumour suppressor *Pml* and revealed that Pml has an indispensable role in maintaining LIC quiescence. *Pml*-deficient LICs become exhausted with time and are incapable of generating CML in transplanted animals. Hence we proposed that there could be a therapeutic window in targeting PML for therapy.

On the basis of this assumption, we used As_2O_3—a drug that downregulates PML expression by targeting it for degradation[33] and is currently used for the treatment of APL with very limited toxicity[30]—to mimic loss of *Pml*. Inhibition of Pml by As_2O_3 disrupted LIC maintenance and increased the efficacy of anti-leukaemic therapy by sensitizing LICs to pro-apoptotic stimuli. Consistent with the notion that targeting the quiescent LICs might be an effective strategy to cure CML, administration of growth factors or bryostatin-1 was previously shown to reduce quiescent CML cells and residual disease after imatinib treatment *ex vivo*[6,37]. Treatment with As_2O_3 in CML might prove therapeutically beneficial because this agent is already in the clinic where it already proved to be extremely well tolerated in extensive preclinical trials in mouse models[33,38,39] and in human APL[30–32].

Finally, although loss of *Pml* and As_2O_3 treatment also induce cycling of HSCs, $Pml^{-/-}$ HSCs are less affected than LICs and can sustain a normal lifespan in the mouse. On the basis of our findings, we therefore propose that As_2O_3 or novel PML-lowering drugs should be used transiently at leukaemia onset, along with, or followed by, a standard of care regimens.

In previous studies, we showed that loss of *Pml* leads to an acceleration of APL in mouse models[40,41]. These data are coherent with the initial acceleration of *Pml*-deficient CML reported in this work (Supplementary Fig. 5a, b). Although there is no clear evidence that APL originates from a HSC, future work will be important to establish if *Pml* deficiency or As_2O_3 treatment leads to the exhaustion of the APL LIC in serial transplantation experiments. This is a plausible hypothesis coherent with the fact that PML–RAR-α may not inhibit PML function completely and that As_2O_3 is extremely effective in the treatment of APL.

Our data demonstrate an unexpected and critical role for PML in stem cell biology and point at its therapeutic targeting as a promising avenue to eradicate LICs in leukaemia. It remains to be determined whether PML exerts a similar role in stem cells in other tissues and in CICs in other tumours, and if the transient use of As_2O_3 may therefore represent a more global strategy to target the CIC in other forms of cancer.

METHODS SUMMARY

Mice. Generation of *Pml*-deficient mice (129Sv) has been described[42]. As_2O_3 (2.5 mg per kg body weight per day) was administered by intraperitoneal injection as described[39].

Western blot. Each lineage compartment was flow-sorted directly into individual wells of a U-bottom 96-well plate containing 2× protein sample buffer. The lysate was briefly boiled and analysed by immunoblotting. The following antibodies were used: anti-β-actin (A-5316, Sigma), anti-mouse Pml (S36 and S37 monoclonal antibodies, provided by S. Lowe) and anti-human PML (Chemicon, rabbit polyclonal antibody) for human PML. Proteins were visualized using the SuperSignal western blotting kit (Pierce). Relative protein expression signals were normalized by comparison with β-actin signals.

Bone-marrow infection and transplantation experiments. Bone-marrow infection and transplantation into lethally irradiated Ly45.1 congenic mice were performed as reported[43]. For serial transplantation, 2.5×10^6 bone marrow mononuclear cells were collected from recipient mice 2 weeks after BMT (first BMT) and transplanted into other recipient mice (second BMT). Subsequent transplantations were performed in the same manner. In some experiments, recipient mice were intraperitoneally injected with As_2O_3.

Primary patient sample assay. Bone marrow samples from healthy volunteers and patients with chronic-phase CML before any therapy at diagnosis were obtained according to appropriate Human Protection Committee validation at the Keio University School of Medicine (Tokyo, Japan) and at the Beth Israel Deaconess Medical Center (Boston, Massachusetts, USA) with written informed consent. Cells were maintained in serum-free medium with a cytokine mixture containing 100 ng ml^{-1} of stem cell factor (SCF), 100 ng ml^{-1} of Flt-3 ligand (Peprotech) and 100 ng ml^{-1} of thrombopoietin (TPO) (Peprotech). To investigate division of HSCs and LICs, sorted $Lin^-CD34^+CD38^{low/neg}$ cells from healthy volunteers and CML patients were stained with CFSE (Molecular Probes). After 3 days of culture, fluorescence intensity of CFSE was analysed by FACS.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 December 2007; accepted 22 April 2008.
Published online 11 May 2008.

1. Bruce, W. R. & van der Gaag, H. A quantitative assay for the number of murine lymphoma cells capable of proliferation *in vivo*. *Nature* **199**, 79–80 (1963).

2. Huntly, B. J. & Gilliland, D. G. Leukaemia stem cells and the evolution of cancer-stem-cell research. *Nature Rev. Cancer* **5,** 311–321 (2005).
3. Scadden, D. T. Cancer stem cells refined. *Nature Immunol.* **5,** 701–703 (2004).
4. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–111 (2001).
5. Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunol.* **5,** 738–743 (2004).
6. Holtz, M., Forman, S. J. & Bhatia, R. Growth factor stimulation reduces residual quiescent chronic myelogenous leukemia progenitors remaining after imatinib treatment. *Cancer Res.* **67,** 1113–1120 (2007).
7. Wang, J. C. & Dick, J. E. Cancer stem cells: lessons from leukemia. *Trends Cell Biol.* **15,** 494–501 (2005).
8. Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–293 (1973).
9. de Klein, A. *et al.* A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* **300,** 765–767 (1982).
10. Deininger, M. W., Goldman, J. M. & Melo, J. V. The molecular biology of chronic myeloid leukemia. *Blood* **96,** 3343–3356 (2000).
11. Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344,** 1031–1037 (2001).
12. Kantarjian, H. *et al.* Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *N. Engl. J. Med.* **346,** 645–652 (2002).
13. Rousselot, P. *et al.* Imatinib mesylate discontinuation in patients with chronic myelogenous leukemia in complete molecular remission for more than 2 years. *Blood* **109,** 58–60 (2007).
14. Ghanima, W., Kahrs, J., Dahl, T. G. III & Tjonnfjord, G. E. Sustained cytogenetic response after discontinuation of imatinib mesylate in a patient with chronic myeloid leukaemia. *Eur. J. Haematol.* **72,** 441–443 (2004).
15. Cortes, J., O'Brien, S. & Kantarjian, H. Discontinuation of imatinib therapy after achieving a molecular response. *Blood* **104,** 2204–2205 (2004).
16. Mauro, M. J., Druker, B. J. & Maziarz, R. T. Divergent clinical outcome in two CML patients who discontinued imatinib therapy after achieving a molecular remission. *Leuk. Res.* **28,** 71–73 (2004).
17. Merante, S. *et al.* Outcome of four patients with chronic myeloid leukemia after imatinib mesylate discontinuation. *Haematologica* **90,** 979–981 (2005).
18. Higashi, T. *et al.* Imatinib mesylate-sensitive blast crisis immediately after discontinuation of imatinib mesylate therapy in chronic myelogenous leukemia: report of two cases. *Am. J. Hematol.* **76,** 275–278 (2004).
19. Bernardi, R. & Pandolfi, P. P. Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nature Rev. Mol. Cell Biol.* **8,** 1006–1016 (2007).
20. Salomoni, P. & Pandolfi, P. P. The role of PML in tumor suppression. *Cell* **108,** 165–170 (2002).
21. Wang, Z. G. *et al.* PML is essential for multiple apoptotic pathways. *Nature Genet.* **20,** 266–272 (1998).
22. Bernardi, R. *et al.* PML inhibits HIF-1α translation and neoangiogenesis through repression of mTOR. *Nature* **442,** 779–785 (2006).
23. Gurrieri, C. *et al.* Loss of the tumor suppressor PML in human cancers of multiple histologic origins. *J. Natl Cancer Inst.* **96,** 269–279 (2004).
24. Scaglioni, P. P. *et al.* A CK2-dependent mechanism for degradation of the PML tumor suppressor. *Cell* **126,** 269–283 (2006).
25. Arai, F. *et al.* Tie2/angiopoietin-1 signaling regulates hematopoietic stem cell quiescence in the bone marrow niche. *Cell* **118,** 149–161 (2004).
26. Pardal, R., Clarke, M. F. & Morrison, S. J. Applying the principles of stem-cell biology to cancer. *Nature Rev. Cancer* **3,** 895–902 (2003).
27. Daley, G. Q., Van Etten, R. A. & Baltimore, D. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. *Science* **247,** 824–830 (1990).
28. Klaassen, C. D. Heavy metals and heavy-metal antagonists. In *The Pharmacological Basis of Therapeutics* (eds Hardman, J. G., Gilman, A. G. & Limbird, L. E.) 1649–1672 (McGraw-Hill, New York, 1996).
29. Aronson, S. M. Arsenic and old myths. *R. I. Med.* **77,** 233–234 (1994).
30. Mathews, V. *et al.* Single-agent arsenic trioxide in the treatment of newly diagnosed acute promyelocytic leukemia: durable remissions with minimal toxicity. *Blood* **107,** 2627–2632 (2006).
31. Soignet, S. L. *et al.* Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide. *N. Engl. J. Med.* **339,** 1341–1348 (1998).
32. Shen, Z. X. *et al.* Use of arsenic trioxide (As_2O_3) in the treatment of acute promyelocytic leukemia (APL): II. Clinical efficacy and pharmacokinetics in relapsed patients. *Blood* **89,** 3354–3360 (1997).
33. Lallemand-Breitenbach, V. *et al.* Role of promyelocytic leukemia (PML) sumolation in nuclear body formation, 11S proteasome recruitment, and As_2O_3-induced PML or PML/retinoic acid receptor α degradation. *J. Exp. Med.* **193,** 1361–1371 (2001).
34. Yilmaz, O. H. *et al.* Pten dependence distinguishes haematopoietic stem cells from leukaemia-initiating cells. *Nature* **441,** 475–482 (2006).
35. Zhang, J. *et al.* PTEN maintains haematopoietic stem cells and acts in lineage choice and leukaemia prevention. *Nature* **441,** 518–522 (2006).
36. Bhatia, R. *et al.* Persistence of malignant hematopoietic progenitors in chronic myelogenous leukemia patients in complete cytogenetic remission following imatinib mesylate treatment. *Blood* **101,** 4701–4707 (2003).
37. Jørgensen, H. G. *et al.* Enhanced CML stem cell elimination *in vitro* by bryostatin priming with imatinib mesylate. *Exp. Hematol.* **33,** 1140–1146 (2005).
38. Lallemand-Breitenbach, V. *et al.* Retinoic acid and arsenic synergize to eradicate leukemic cells in a mouse model of acute promyelocytic leukemia. *J. Exp. Med.* **189,** 1043–1052 (1999).
39. Rego, E. M., He, L. Z., Warrell, R. P. Jr, Wang, Z. G. & Pandolfi, P. P. Retinoic acid (RA) and As_2O_3 treatment in transgenic models of acute promyelocytic leukemia (APL) unravel the distinct nature of the leukemogenic process induced by the PML-RARα and PLZF-RARα oncoproteins. *Proc. Natl Acad. Sci. USA* **97,** 10173–10178 (2000).
40. Rego, E. M. *et al.* Role of promyelocytic leukemia (PML) protein in tumor suppression. *J. Exp. Med.* **193,** 521–529 (2001).
41. Gurrieri, C. *et al.* Mutations of the PML tumor suppressor gene in acute promyelocytic leukemia. *Blood* **103,** 2358–2362 (2004).
42. Wang, Z. G. *et al.* Role of PML in cell growth and the retinoic acid pathway. *Science* **279,** 1547–1551 (1998).
43. Di Cristofano, A. *et al.* $p62^{dok}$, a negative regulator of Ras and mitogen-activated protein kinase (MAPK) activity, opposes leukemogenesis by $p210^{bcr\text{-}abl}$. *J. Exp. Med.* **194,** 275–284 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Ogilvie for analysis of patient samples and data management, and all members of the Pandolfi laboratory for comments and discussion. K.I. was supported by a JSPS postdoctoral fellowship for research abroad. R.B. is supported by a K01 NIH grant. This work was supported by NIH grants to P.P.P.

Author Contributions The experiments were conceived and designed by K.I., R.B., A.M. and P.P.P. Experiments were performed by K.I., R.B. and A.M. Immunohistochemistry of patient samples was conducted and investigated by J.T.-F, K.I. and R.B. Experiments on primary human CML samples were performed by A.M., S.M., G.S., Y.I., J.R., K.I. and D.E.A. Data were analysed by K.I., R.B., A.M. and P.P.P. The paper was written by K.I., R.B., A.M. and P.P.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.P.P. (ppandolf@bidmc.harvard.edu).

METHODS

Mice. Generation of *Pml*-deficient mice (129Sv) has been described[42]. C57BL/6 mice (B6-CD45.2) and C57BL/6 mice congenic for the CD45 locus (B6-CD45.1) were purchased from The Jackson Laboratory and crossed with 129Sv mice. F_3 B6/Sv129 mice were used as recipients in transplantation assays. As_2O_3 (2.5 mg per kg body weight per day) was administered by intraperitoneal injection as described[39].

Long-term cultures and colony-forming assays. For long-term cultures, KSL cells were co-cultured with stromal cells in minimum essential media α modification (αMEM, Sigma) containing 12.5% FCS (JRH Bioscience), 12.5% horse serum (Gibco BRL) and 1.0 nM dexamethasone. After 2, 4 or 6 weeks of culture, cells were harvested and used for haematopoietic colony-forming assays as described[44]. For some experiments, 0.15 μM As_2O_3 (Sigma) or 0.10 μM Ara-C (Sigma) was added to cultures.

Western blot. Each lineage compartment was flow-sorted directly into individual wells of a U-bottom 96-well plate containing 50 μl 2× protein sample buffer. The lysate was briefly boiled and analysed by immunoblotting. The following antibodies were used: anti-β-actin (A-5316, Sigma), anti-mouse Pml (S36 and S37 monoclonal antibodies, provided by S. Lowe) and anti-human PML (Chemicon, rabbit polyclonal antibody) for human PML, anti-PS6 (S235/236), and anti-S6 (Cell Signaling). Proteins were visualized using the SuperSignal western blotting kit (Pierce). Signal intensity was measured using ImageJ 1.34S software (http://rsb.info.nih.gov/ij/). Relative protein expression signals were normalized by comparison with β-actin signals.

Bone marrow infection and transplantation experiments. Transfection of the retroviral vector $p210^{BCR\text{-}ABL}$, bone marrow isolation from 8-week-old wild-type and $Pml^{-/-}$ mice, pre-stimulation and infection, and transplantation into lethally irradiated Ly45.1 congenic mice were performed as reported[43]. For serial transplantation, 2.5×10^6 bone marrow mononuclear cells were collected from recipient mice 2 weeks after BMT (first BMT). Collected cells were transplanted into other recipient mice (second BMT). Subsequent transplantations were performed in the same manner. In some experiments, recipient mice were intraperitoneally injected with As_2O_3. For minimal residual disease, a two-round nested PCR reaction was applied as described[45].

Primary patient sample assay. Bone marrow samples from healthy volunteers (three males and one female, median age 31.5 years: range 29–36 years) and patients with chronic-phase CML before any therapy at diagnosis (two males and two females, median age at diagnosis 32.5 years: range 29–37 years; percentage blasts, 1.5 to 3.8; cytogenetics at diagnosis, Philadelphia chromosome (Ph1) was detected in all patients) were obtained according to appropriate Human Protection Committee validation at the Keio University School of Medicine (Tokyo, Japan) and at the Beth Israel Deaconess Medical Center (Boston, Massachusetts, USA) with written informed consent. Mononuclear cells were separated by Lymphoprep (Nycomed Pharma As). Cells were maintained in serum-free medium with a cytokine mixture containing 100 ng ml^{-1} of stem cell factor (SCF), 100 ng ml^{-1} of Flt-3 ligand (Peprotech) and 100 ng ml^{-1} of thrombopoietin (TPO) (Peprotech). For some experiments, 0.15 μM As_2O_3 (Sigma) (day 1–7) or 0.10 μM Ara-C (Sigma) (day 3–7) was added to cultures. To investigate division of HSCs and LICs, sorted $Lin^-CD34^+CD38^{low/neg}$ cells from healthy volunteers and CML patients were stained with CFSE (Molecular Probes). CFSE-stained cells were plated on 96-well U-bottom plates. After 3 days of culture, fluorescence intensity of CFSE was analysed by FACS. To assay cell division at the single cell level, single cell sorting of $Lin^-CD34^+CD38^{low/neg}c\text{-}Kit^+$ cells was performed and cells were cultured with SCF plus TPO plus Flt-3L. Cell division was monitored daily.

Statistical analysis. *P*-values were calculated using the unpaired Student's *t*-test.

44. Ito, K. *et al.* Reactive oxygen species act through p38 MAPK to limit the lifespan of hematopoietic stem cells. *Nature Med.* **12**, 446–451 (2006).
45. Cross, N. C. *et al.* Minimal residual disease after allogeneic bone marrow transplantation for chronic myeloid leukaemia in first chronic phase: correlations with acute graft-versus-host disease and relapse. *Br. J. Haematol.* **84**, 67–74 (1993).

LETTERS

Surprising dissimilarities in a newly formed pair of 'identical twin' stars

Keivan G. Stassun[1], Robert D. Mathieu[2], Phillip A. Cargile[1], Alicia N. Aarnio[1], Eric Stempels[3] & Aaron Geller[2]

The mass and chemical composition of a star are the primary determinants of its basic physical properties—radius, temperature and luminosity—and how those properties evolve with time[1]. Accordingly, two stars born at the same time, from the same natal material and with the same mass, are 'identical twins,' and as such might be expected to possess identical physical attributes. We have discovered in the Orion nebula a pair of stellar twins in a newborn binary star system[2]. Each star in the binary has a mass of 0.41 ± 0.01 solar masses, identical to within 2 per cent. Here we report that these twin stars have surface temperatures differing by ~300 K (~10 per cent) and luminosities differing by ~50 per cent, both at high confidence level. Preliminary results indicate that the stars' radii also differ, by 5–10 per cent. These surprising dissimilarities suggest that one of the twins may have been delayed by several hundred thousand years in its formation relative to its sibling. Such a delay could only have been detected in a very young, definitively equal-mass binary system[3]. Our findings reveal cosmic limits on the age synchronization of young binary stars, often used as tests for the age calibrations of star-formation models[4].

Astronomers have long relied on eclipsing binary star systems—in which two stars periodically eclipse one another as they orbit—to measure the basic physical properties of stars and for testing the most fundamental predictions of theoretical stellar evolution models[1]. Because the two stars in a binary system are presumed to have formed at the same time and from the same parent cloud material, eclipsing binaries permit a direct test of theoretical models with mass as the primary independent variable. In recent years the discovery and analysis of eclipsing binary systems have been fruitfully applied to probe the basic physical properties of newborn low-mass stars and brown dwarfs that are still in the earliest stages of evolution[5–9].

Par 1802 is a young (age of ~1 Myr) eclipsing binary system in the Orion nebula cluster recently discovered by us[2], and is the youngest equal-mass eclipsing binary so far found. It is therefore a particularly sensitive case study with which to test to what extent stars of the same mass, age and composition are identical, independently of theoretical models. In fact, we find that the components of Par 1802 possess clearly dissimilar surface temperatures and luminosities, and probably dissimilar radii as well.

From measurements of the radial velocities of Par 1802 obtained with the Hobby–Eberly Telescope's high-resolution spectrograph, we have previously determined the orbital parameters of the system[2], including a binary mass ratio of $q = 1.03 \pm 0.03$. With the addition of a precise light curve obtained from the 0.9-m telescope at the Kitt Peak National Observatory and the SMARTS telescopes at the Cerro Tololo Inter-American Observatory (Fig. 1), we are now able to perform a combined analysis[10,11] of the light-curve and radial-velocity data, which yields accurate measurements of the fundamental orbital and physical parameters of Par 1802 (Table 1). The refined orbit solution gives an improved mass ratio of $q = 0.98 \pm 0.01$. Thus, the components of Par 1802 are twins with masses of 0.41 ± 0.01 solar masses ($M_\odot$) in an orbit that is very nearly circular.

The relative depths of the eclipses (Fig. 1) yield the ratio of the stars' surface brightnesses at a wavelength of 0.8 μm, from which we determine the ratio of their surface temperatures to be $T_1/T_2 = 1.085 \pm 0.007$. Our light-curve analysis includes up-to-date model

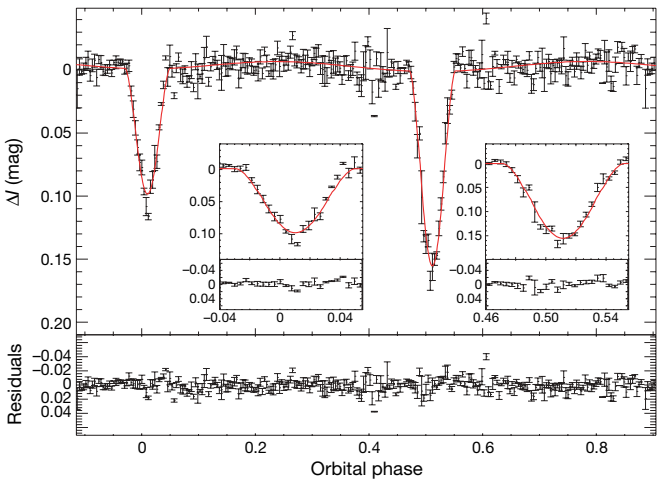


Figure 1 | Light curve of Par 1802 at I band (0.8 μm). We repeatedly imaged Par 1802 with the 0.9-m telescope at the Kitt Peak National Observatory and with the SMARTS 0.9-m, 1.0-m and 1.3-m telescopes at the Cerro Tololo Inter-American Observatory, from December 1994 to March 2007. In total, 2,209 flux measurements were obtained on 418 separate nights and with an average cadence of five or six measurements per night. The typical relative uncertainty in the individual flux measurements is ~1%. A phase dispersion minimization analysis[21] reveals an unambiguous period of $P = 4.673843 \pm 0.000068$ days. The measurements are shown here as differential magnitudes ΔI, folded on the above period and phased relative to periastron passage (that is, closest approach of the two stars to one another), as determined from the orbit solution (Table 1). The measurements have been resampled into 250 bins equally spaced in phase; each data point plotted represents the average of about nine measurements, and error bars represent the root mean square of the values that were averaged together. Note that the fitting procedure (see Table 1) was performed on the raw data, not the resampled light curve shown here for visual clarity. The ratio of eclipse depths provides a direct measure of the ratio of surface temperatures, with the deeper eclipse corresponding to the eclipse of the hotter component. As a result of the eccentricity and orientation of periastron, the eclipse of the primary component occurs near orbital phase 0.52. A model fit to the light curve incorporating the orbital and physical parameters of Par 1802 (Table 1) is also shown (solid red curve). Insets show the eclipses in detail, and the lower panel shows the residuals in magnitudes of the data relative to the model (observed minus calculated).

[1]Department of Physics and Astronomy, Vanderbilt University, Nashville, Tennessee 37235, USA. [2]Department of Astronomy, University of Wisconsin—Madison, Madison, Wisconsin 53706, USA. [3]School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK.

flux spectra for low-mass stars[12], although the simple fact of unequal eclipse depths in a system with a nearly circular orbit by itself makes the finding of unequal surface temperatures incontrovertible.

Par 1802 has previously[13] been assigned a spectral type of M2 with an uncertainty of ±1 subtype, corresponding to a temperature of ~3,560 ± 150 K (ref. 4). Similarly, fitting a single-temperature model stellar atmosphere[12] to the observed fluxes of Par 1802 from 0.35 μm to 8 μm (Supplementary Table 1) gives a temperature of 3,800 ± 100 K. This, together with the temperature ratio above, implies component temperatures of $T_1 = 3{,}945 \pm 15$ K and $T_2 = 3{,}655 \pm 15$ K, with the same (correlated) systematic uncertainty of 100 K in both.

In addition, from the observed eclipse durations and orbital velocities we measure the sum of the stars' radii to be $R_1 + R_2 = 3.51 \pm 0.05$ solar radii ($R_\odot$). At the precision of our light-curve data (~1%), the nearly circular orbit precludes a determination of the individual radii from the single-band light curve alone; in a circular system, the primary and secondary eclipse durations are identical, and thus any combination of stellar radii that sum to the same value can reproduce the observed light curve at the level of ~0.1%. Light-curve observations at multiple wavelengths will help to remove this degeneracy in the component radii.

In the meantime, the radii can be estimated from the flux ratio of the two stars. We have performed a spectral decomposition analysis[14,15] of the component spectra of Par 1802 (see Supplementary Information), from which we find $F_2/F_1 = 0.55 \pm 0.06$ at 0.6 μm. This, combined with the temperature ratio and appropriate bolometric corrections[16], implies a radius ratio of $R_1/R_2 = 1.08 \pm 0.05$, or $R_1 = 1.82 \pm 0.05 R_\odot$ and $R_2 = 1.69 \pm 0.05 R_\odot$. This difference of 5–10% in the radii is consistent with the observed projected rotational velocities (Table 1) if the stars are rotating synchronously at the orbital period with spin axes parallel to the orbital axis. Finally, the ratio of stellar luminosities that results, through the Stefan–Boltzmann law, from the ratios of temperatures and radii is $L_1/L_2 = 1.58 \pm 0.10$, which gives $L_1 = 0.72 \pm 0.11\ L_\odot$ and $L_2 = 0.46 \pm 0.12\ L_\odot$.

Table 1 | Orbital and physical parameters of Par 1802

Parameter	Value
Time of periastron passage	2003.834996 ± 0.000055
Eccentricity, e	0.029 ± 0.005
Orientation of periastron, ω	266.1 ± 1.8°
Semimajor axis, $a\sin i$	0.0501 ± 0.0006 AU
Centre-of-mass velocity, γ	23.7 ± 0.5 km s^{-1}
Mass ratio, $q \equiv M_2/M_1$	0.98 ± 0.01
Total mass, $(M_1 + M_2)\sin^3 i$	0.768 ± 0.028$M_\odot$
Inclination, i	78.1 ± 0.6°
Primary mass, M_1	0.414 ± 0.015$M_\odot$
Secondary mass, M_2	0.406 ± 0.014$M_\odot$
Sum of radii, $R_1 + R_2$	3.51 ± 0.05$R_\odot$
Surface temperature ratio, T_1/T_2	1.084 ± 0.007
Primary rotation speed, $v_1 \sin i$	17 ± 2 km s^{-1}
Secondary rotation speed, $v_2 \sin i$	14 ± 3 km s^{-1}

We simultaneously fitted the radial-velocity data from ref. 2 and the light curve in Fig. 1 with a standard detached-eclipsing-binary model[10,11]. The code assumes full Roche geometry according to the formalism of Kopal[10], and includes model atmospheres[12] to determine intensities over the stellar discs. We adopted a linear limb-darkening law and allowed the code to calculate reflection effects, adopting a bolometric albedo of 0.5, typical for fully convective stellar atmospheres. To maintain control of the solution and its many parameters, we performed this fitting in stages. In the initial stage we fixed the orbital parameters at the values from ref. 2, and assumed an average surface temperature of 3,800 K for the components (see the text) and solar metallicity. This allowed us to obtain initial estimates of the component temperatures and the sum of the radii, and the system inclination. With these initial values so determined, we then iteratively improved the solution by first fitting the eccentricity and orientation of periastron, and then performing a final fit in which all of the orbital and component parameters of the system were fitted freely. Uncertainties in the parameters represent standard 1σ (s.e.m.) formal errors from the covariance matrix of the eclipsing-binary model fit[10,11]. The most important degeneracy in the solution is between the component radii; because of the nearly circular orbit, almost any combination of radii that sum to the same total value will fit the light curve equally well. We have measured the projected rotation speeds, $v\sin i$, of the two components by comparing the widths of the observed cross-correlation peaks[2] with those of an M2 star whose spectrum was rotationally broadened artificially. The $v\sin i$ values and uncertainties are the averages and standard deviations resulting from five observations of Par 1802 near maximum radial-velocity separation of the two components[2]. Note that both here and in ref. 2 the more luminous star (here the formally more massive star) is identified as the 'primary.'

We refitted the observed spectral energy distribution of Par 1802 from 0.35 μm to 8 μm (Supplementary Table 1) by using synthetic spectra[12] of cool stars with the above temperatures and radii (Fig. 2). Only the distance and reddening due to extinction were varied in the fit. The observed spectral energy distribution is very well fitted with an extinction of $A_v = 0.5 \pm 0.2$ visual magnitudes and a distance of 420 ± 15 pc, which is consistent with the low-mass stellar population associated with the Orion nebula at 480 ± 80 pc (ref. 17). This star-forming region is very young, with an estimated age of 1^{+2}_{-1} Myr (ref. 13).

There is weak evidence for an excess of infrared emission at wavelengths longer than 5 μm (Fig. 2), perhaps indicative of a circumbinary disc and/or a faint tertiary body. In this light, the slight eccentricity of the orbit (Table 1) might also be taken as evidence for a tertiary body in the system. Flux measurements at longer wavelengths will be required for verification of this possibility.

Unequal temperatures and luminosities for the equal-mass stars of Par 1802 are securely established in the current analysis, and unequal radii are also suggested. Some stellar evolution models for young low-mass stars[3] predict that stars with masses of 0.4$M_\odot$ undergo a brief period of rapid evolution at an age of ~1 Myr (Fig. 3). In particular, the observed temperatures, luminosities and radii of Par 1802 are consistent with the model predictions if the warmer, larger, more luminous star is interpreted as being slightly less evolved than its companion. Because of the predicted rapidity of evolution, a 'lag' of only a few hundred thousand years would be required (Fig. 3).

Such an age difference is only observationally detectable in an equal-mass binary system during the first few million years of its evolution, where the models predict that changes in the stars' physical properties are fastest and most pronounced. At later times any physical signs of non-synchronized ages will have become smaller than 1%, below the precision limit of the best observations (Fig. 3), and in a young binary with components of unequal masses, uncertain theoretical models will convert precise physical differences into an imprecise age difference. That Par 1802—the first very young, definitively equal-mass binary to be studied—shows evidence for age differences between its stars suggests that this may be a common feature of young binaries.

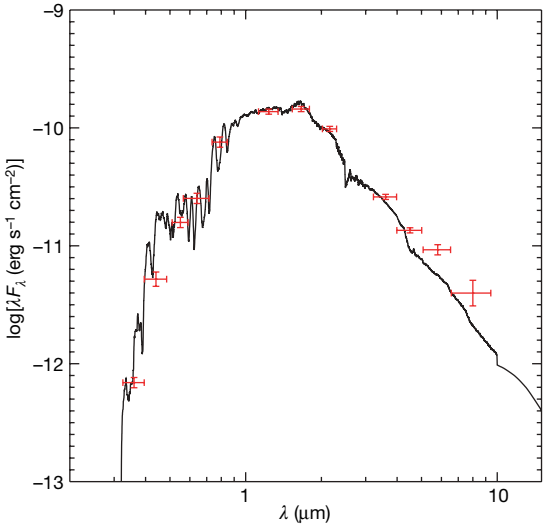


Figure 2 | Spectral energy distribution of Par 1802. Broadband flux measurements of Par 1802 from 0.35 μm to 8 μm (Supplementary Table 1; refs 22–30) are shown in red. Vertical error bars represent s.e.m. uncertainties in the flux measurements; horizontal bars represent the filter bandpasses used for the flux measurements. The solid curve is a composite of two synthetic spectra[12] of young, low-mass stars with temperatures, masses and radii corresponding to those measured for the components of Par 1802 (Table 1). Fitting for extinction and distance, we find $A_v = 0.5 \pm 0.2$ mag and $d = 420 \pm 15$ pc.

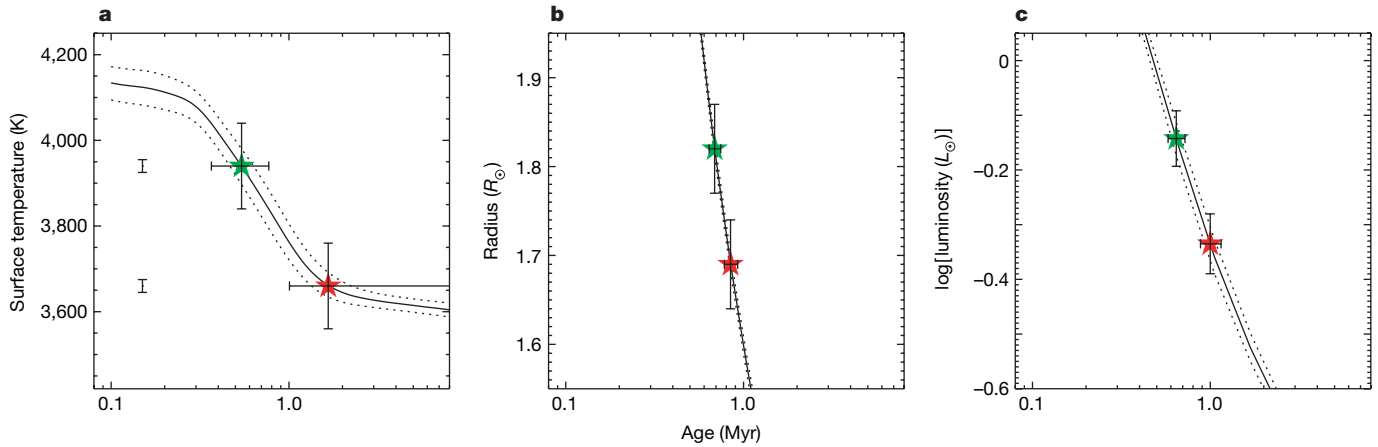


Figure 3 | Comparison of physical properties of Par 1802 with theoretical predictions. a, Surface temperature; **b**, radius; **c**, luminosity. In each panel, the solid line shows the predicted evolution of a $0.41M_\odot$ star with solar composition from the theoretical models of ref. 3. Dotted lines show the result of changing the stellar mass by $\pm 0.015M_\odot$, representative of the uncertainties in the measured masses of Par 1802 (Table 1). The measured properties of the primary and secondary components of Par 1802 are shown as green and red symbols, respectively. The small vertical bars at the left of **a** represent the measurement errors alone (that is, not including the ~100-K systematic uncertainty in the absolute temperature scale) resulting from the precisely measured temperature ratio. Vertical error bars on the points represent the combination of measurement and systematic uncertainties (see the text). Horizontal error bars represent the range of ages for which the theoretical models are consistent with the measurements within the uncertainties (including systematic uncertainties). The stellar luminosities plotted in **c** are calculated from the measured radii and surface temperatures. Note that the uncertainties in the temperatures, radii and luminosities are not independent between the two stars, because they are connected by precisely determined ratios; thus, for example, the primary star cannot be forced cooler while simultaneously forcing the secondary warmer. The nominal age of the Orion nebula cluster is ~1 Myr (ref. 13).

Current theories of binary formation are largely silent on the formation of such short-period binaries, and so we have no theoretical context within which to interpret such a difference in evolutionary age. Certainly the lack of synchronization of the stellar evolution clocks will provide a new insight into the formation processes of short-period binaries.

Alternatively, the stars may be the same age but have differing properties despite their very similar masses and (presumably) composition. There is mounting evidence that strong magnetic fields on the surfaces of young stars may alter their physical properties[8,18–20]. Thus, for example, one of the stars in Par 1802 may possess a strong magnetic field that is also substantially different in strength or geometry from that of its companion. However, we have observed that the strength of the Balmer α line of hydrogen, a commonly used tracer of magnetic activity in low-mass stars, is very weak in both components of Par 1802 (at most a few hundred milliångströms of emission)[2]. Moreover, the observational and theoretical evidence so far obtained for the effects of magnetic fields on stellar properties indicates that only the surface temperatures and radii of the stars should be affected. The luminosities of the stars are not predicted to be significantly modified by magnetic fields, because the luminosities are primarily determined by internal processes[19]. These expectations are at odds with the finding of a factor of ~2 luminosity difference between the stars in Par 1802. In any case, even if it were correct, this hypothesis would still leave unanswered the question of why these two stars with essentially identical masses and rotation rates do not possess similar magnetic properties.

Finally, young binary systems have been used as test beds of the ories of early stellar evolution. The agreement of theoretical ages derived for each star in a binary system is taken as a self-consistency test for pre-main-sequence stellar evolution models[4]. The lack of age synchronization in Par 1802 suggests a precision limit of several hundred thousand years for such empirical tests.

Par 1802 provides direct evidence that birth order in 'identical twin' stars can manifest itself as observable physical differences between the two stars—at least when they are very young.

Received 7 February; accepted 2 May 2008.

1. Andersen, J. Accurate masses and radii of normal stars. *Astron. Astrophys. Rev.* **3**, 91–126 (1991).
2. Cargile, P. A., Stassun, K. G. & Mathieu, R. D. Discovery of Par 1802 as a low-mass pre-main-sequence eclipsing binary in the Orion star-forming region. *Astrophys. J.* **674**, 329–335 (2008).
3. D'Antona, F. & Mazzitelli, I. Evolution of low mass stars. *Mem. Soc. Astron. It.* **68**, 807–822 (1997).
4. Luhman, K. L. Young low-mass stars and brown dwarfs in IC 348. *Astrophys. J.* **525**, 466–481 (1999).
5. Stassun, K. G., Mathieu, R. D., Vaz, L. P. R., Stroud, N. & Vrba, F. J. Dynamical mass constraints on low-mass pre-main-sequence stellar evolutionary tracks: an eclipsing binary in Orion with a 1.0 $M_\odot$ primary and a 0.7 $M_\odot$ secondary. *Astrophys. J.* **151** (Suppl.), 357–385 (2004).
6. Covino, E., Frasca, A., Alcala, J. M., Paladino, R. & Sterzik, M. F. Improved fundamental parameters for the low-mass pre-main sequence eclipsing system RX J0529.4 + 0041. *Astron. Astrophys.* **427**, 637–649 (2004).
7. Stassun, K. G., Mathieu, R. D. & Valenti, J. Discovery of two young brown dwarfs in an eclipsing binary system. *Nature* **440**, 311–314 (2006).
8. Stassun, K. G., Mathieu, R. D. & Valenti, J. A surprising reversal of temperatures in the brown dwarf eclipsing binary 2MASS J05352184–0546085. *Astrophys. J.* **664**, 1154–1166 (2007).
9. Irwin, J. *et al.* The Monitor project: JW 380: a 0.26-, 0.15-$M_\odot$, pre-main-sequence eclipsing binary in the Orion nebula cluster. *Mon. Not. R. Astron. Soc.* **380**, 541–550 (2007).
10. Wilson, R. E. & Devinney, E. J. Realization of accurate close-binary light curves: application to MR Cygni. *Astrophys. J.* **166**, 605–620 (1971).
11. Prsa, A. & Zwitter, T. A computational guide to physics of eclipsing binaries. I. Demonstrations and perspectives. *Astrophys. J.* **628**, 426–438 (2005).
12. Allard, F., Hauschildt, P. H. & Schweitzer, A. Spherically symmetric model atmospheres for low-mass pre-main-sequence stars with effective temperatures between 2000 and 6800 K. *Astrophys. J.* **539**, 366–371 (2000).
13. Hillenbrand, L. A. On the stellar population and star-forming history of the Orion Nebula Cluster. *Astron. J.* **113**, 1733–1768 (1997).
14. Bagnuolo, W. G. & Gies, D. R. Tomographic separation of composite spectra—The components of the O-star spectroscopic binary AO Cassiopeiae. *Astrophys. J.* **376**, 266–271 (1991).
15. Zucker, S. & Mazeh, T. Study of spectroscopic binaries with TODCOR. 1: A new two-dimensional correlation algorithm to derive the radial velocities of the two components. *Astrophys. J.* **420**, 806–810 (1994).
16. Slesnick, C. L., Hillenbrand, L. A. & Carpenter, J. M. The spectroscopically determined substellar mass function of the Orion Nebula cluster. *Astrophys. J.* **610**, 1045–1063 (2004).
17. Genzel, R. & Stutzki, J. The Orion molecular cloud and star-forming region. *Annu. Rev. Astron. Astrophys.* **27**, 41–85 (1989).
18. Torres, G., Lacy, C. H., Marschall, L. A., Sheets, H. A. & Mader, J. A. The eclipsing binary V1061 Cygni: Confronting stellar evolution models for active and inactive solar-type stars. *Astrophys. J.* **640**, 1018–1038 (2006).
19. Chabrier, G., Gallardo, J. & Baraffe, I. Evolution of low-mass star and brown dwarf eclipsing binaries. *Astron. Astrophys.* **472**, L17–L20 (2007).

20. Ribas, I. *et al.* Fundamental properties of low-mass stars. *Mem. Soc. Astron. It.* **79**, *Proc. Workshop 'XXI Century Challenges for Stellar Evolution'* (eds Cassisi, S. & Salaris, M.) (in the press).
21. Stellingwerf, R. F. Period determination using phase dispersion minimization. *Astrophys. J.* **224,** 953–960 (1978).
22. Robberto, M. *et al.* An overview of the HST Treasury Program on the Orion Nebula. *Bull. Am. Astron. Soc.* **37,** 1404 (2005).
23. Zacharias, N. *et al.* The Naval Observatory Merged Astrometric Dataset (NOMAD). *Bull. Am. Astron. Soc.* **36,** 1418 (2004).
24. Skrutskie, M. F. *et al.* The Two Micron All Sky Survey (2MASS). *Astron. J.* **131,** 1163–1183 (2006).
25. Cox, A. N. (ed.) *Allen's Astrophysical Quantities* 4th edn (Springer, New York, 2001).
26. Campins, H., Rieke, G. H. & Lebofsky, M. J. Absolute calibration of photometry at 1 through 5 microns. *Astrophys. J.* **90,** 896–899 (1985).
27. Cousins, A. J. VRI standards in the E regions. *Mem. R. Astron. Soc.* **81,** 25 (1976).
28. Bessell, M. S. UBVRI photometry. II—The Cousins VRI system, its temperature and absolute flux calibration, and relevance for two-dimensional photometry. *Publ. Astron. Soc. Pacif.* **91,** 589–607 (1979).
29. Cohen, M., Wheaton, W. A. & Megeath, S. T. Spectral irradiance calibration in the infrared. XIV. The absolute calibration of 2MASS. *Astron. J.* **126,** 1090–1096 (2003).
30. Fazio, G. An IRAC Survey of the L1630 and L1641 (Orion) Molecular Clouds. *Spitzer Space Telescope - Guaranteed Time Observer Proposal #43* ⟨http://ssc.spitzer.caltech.edu/geninfo/gto/abs/43.txt⟩.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Prsa for software used in our analyses. This work is supported by grants to K.G.S. and R.D.M. from the National Science Foundation, and a Cottrell Scholar award to K.G.S. from the Research Corporation. K.G.S. acknowledges the hospitality of the Space Telescope Science Institute's Caroline Herschel Distinguished Visitor programme.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.G.S. (keivan.stassun@vanderbilt.edu).

LETTERS

Jovian-like aurorae on Saturn

Tom Stallard[1], Steve Miller[2], Henrik Melin[3], Makenzie Lystrup[2], Stan W. H. Cowley[1], Emma J. Bunce[1], Nicholas Achilleos[2] & Michele Dougherty[4]

Planetary aurorae are formed by energetic charged particles streaming along the planet's magnetic field lines into the upper atmosphere from the surrounding space environment. Earth's main auroral oval is formed through interactions with the solar wind[1], whereas that at Jupiter is formed through interactions with plasma from the moon Io inside its magnetic field (although other processes form aurorae at both planets[2,3]). At Saturn, only the main auroral oval has previously been observed and there remains much debate over its origin. Here we report the discovery of a secondary oval at Saturn that is ~25 per cent as bright as the main oval, and we show this to be caused by interaction with the middle magnetosphere around the planet. This is a weak equivalent of Jupiter's main oval, its relative dimness being due to the lack of as large a source of ions as Jupiter's volcanic moon Io. This result suggests that differences seen in the auroral emissions from Saturn and Jupiter are due to scaling differences in the conditions at each of these two planets, whereas the underlying formation processes are the same.

There are three competing theories describing the process by which Saturn's main auroral oval is formed: first, that the oval is like that of the Earth, mapping to the boundary between closed field lines and those open to the solar wind, where a shear in rotational flow is expected, requiring a ring of upward-directed current[4]; second, that the oval is associated with centrifugal instabilities in the outer magnetosphere, where variations in solar wind dynamic pressure lead to varying angular velocities, driving currents into the ionosphere[5]; and third, that the oval is an analogue of that at Jupiter, formed by internal magnetospheric processes driven by the rapid rotation of the planet[2,3]. Jupiter's equatorial plasma sheet initially co-rotates with the planet, but as the plasma diffuses away this co-rotation breaks down, resulting in a strong circuit of electric currents. This forces electrons to precipitate along the magnetic field lines into the atmosphere, forming an auroral oval at the jovian latitude mapping to the co-rotation breakdown.

The morphology of Saturn's aurorae has been examined in detail with the use of Hubble Space Telescope images of the ultraviolet emission from the planet[6–8], showing that the aurorae are strongly influenced by changes in the dynamic pressure of the solar wind[9]. Ground-based infrared spectroscopic studies have used emission from the molecular ion H_3^+ to observe Saturn's auroral structure[10] and to measure the associated temperature[11] and ion winds[12] within the upper atmosphere. Recent measurements of the ion wind have shown that the main auroral oval is located significantly poleward of the latitude at which ions no longer co-rotate with the planet, which means that, by definition, Saturn's main auroral oval cannot be Jupiter-like in formation[13]. However, an excess of emission equatorward of the main auroral oval has been observed in a significant proportion of the infrared observations; this aurora could have an analogous origin to that of the main auroral oval on Jupiter.

To test this, it is necessary to isolate this secondary emission from that of the main auroral oval. Long-slit spectrometer measurements cutting perpendicularly through Saturn's auroral region result in contemporaneous profiles of the infrared intensity and corresponding line-of-sight ion velocity (Figs 1 and 2)[14]. Whereas the ultraviolet morphology is dominated by the emission from the main auroral oval, the infrared emission has significant amounts of emission that are not associated with this oval. Assuming that the infrared and ultraviolet main auroral ovals have the same morphological structure[13], a model of the main oval can be based on a statistical analysis of the typical ultraviolet oval[15]. By subtracting this modelled main auroral oval from the infrared intensity profile, the residual emission can be calculated (Fig. 3). This results in three clear regions of secondary auroral emission, one poleward and two equatorward from the location of the main oval.

The relative excess of infrared emission in Saturn's polar cap has previously been noted[14]. This is likely to be caused by significant emission across the polar region, analogous with that of Jupiter, where emission occurs across the entire polar region in both the infrared and ultraviolet[16,17]. Corresponding ultraviolet emission across Saturn's pole, if it exists, is too weak to have been detected.

Of the two intensity peaks equatorward of the main auroral oval, the peak on the dawn side has an intensity ~25% of the main auroral oval and the peak on the dusk side has an intensity ~18% of the main auroral oval. Both these intensity peaks have their maximum values positioned at exactly the same location as the breakdown in co-rotation, measured in the corresponding velocity profile, with the dusk emission centred on the breakdown in co-rotation and the dawn emission somewhat extended, decreasing gently towards the pole. The effect of changing the modelled main oval position on the secondary oval brightness is discussed in more detail in the Supplementary Information.

This correspondence between the position of the peak secondary emission and the breakdown in co-rotation in the ionosphere is direct evidence that the breakdown in co-rotation within the magnetospheric plasma is driving a current system strong enough to produce an H_3^+ aurora, a weaker variant of the main auroral oval seen on Jupiter. This newly identified aurora forms in two distinct regions within the slit, one on each of the dawn and dusk sides, at slightly lower latitudes than the main auroral oval. Given that this flank emission is seen repeatedly in successive runs[13,14] and that there is significant emission equatorward of the main oval at noon[10], this constitutes strong evidence of the presence of a second auroral oval. Auroral electron beams have been detected fairly deep inside the magnetosphere, supporting an internally driven auroral component[18].

These are the first auroral emissions from Saturn that can be directly related to the main auroral oval at Jupiter. So far there has been no published identification of such emission within the ultraviolet data set, the closest comparison being a limb-brightened

[1]Department of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, UK. [2]Atmospheric Physics Laboratory, Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK. [3]Space Environment Technologies, Planetary and Space Science Division, 320 N. Halstead Street, Suite 110, Pasadena, California 91107, USA. [4]Space and Atmospheric Physics Group, Department of Physics, Imperial College of Science, Technology and Medicine, London SW7 2BW, UK.

'auroral zone O' ('O' for 'outer')[19]. The potential for such oval emission was previously predicted, for ultraviolet emission, to form a dim narrow oval too weak (~1 kR and ~1° wide) to be detected with current observation techniques[20]. Saturn's main auroral oval is typically 20–50 kR, so a secondary oval at 20% brightness is at about the limit of detectability with Hubble. This suggests that the strength of the current formed by breakdown of co-rotation at Saturn is stronger than has been previously modelled and that future observations, either in the infrared or in the ultraviolet, should be better able to determine the exact location of the secondary oval, allowing these models to be improved.

With this discovery, our understanding is that Saturn's aurorae are more akin to those of Jupiter. Observations of Jupiter have previously shown that a significant region of solar wind control exists poleward of the main auroral oval[21], a region where there is significant emission in both the infrared and ultraviolet[17,22]. It therefore seems likely that Jupiter's polar aurora is at least partly 'saturnian-like' in origin; if this is so, a secondary oval will be located within the polar region.

Thus, although the general morphology of the aurorae of the planets is significantly different, the processes by which they are formed seem to be similar. Specific morphological differences, such as the strong dawn brightening seen at Saturn during solar wind compressions[8], still distinguish the aurorae from those at Jupiter and have much to tell us about the interaction of the planets with the solar

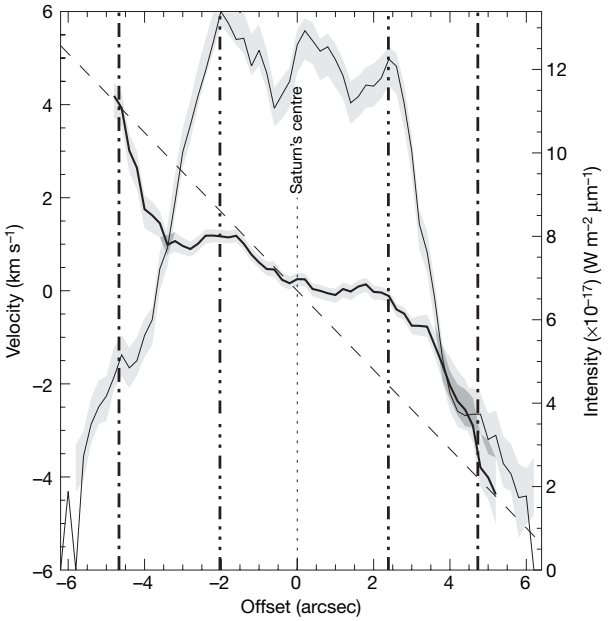


Figure 1 | H_3^+ aurora emission and the associated line-of-sight velocity. The intensity profile (thin) is plotted with the velocity in the inertial frame (bold) and, for reference, the rotation of the planet (dashed). Two vertical dot–dashed lines mark the location where the ionosphere begins to sub-rotate from the neutral atmosphere. Two vertical three-dot–dashed lines mark the location of the main auroral oval within the slit. The edge of the plot delineates the limb of the planet. This data was taken on 2003 February 6 as a part of an extensive set of observations of the auroral/polar regions of Saturn, with the NASA Infrared Telescope Facility. The long-slit spectrometer CSHELL[23] was aligned west–east on the planet, with the field of view of the slit crossing the auroral region. The resultant high-resolution spectrum was centred on emission from the H_3^+ ν_2 Q(1,0$^-$) line at 3.953 μm. This line was fitted with a gaussian at each spatial position, and the gaussian peak brightness and position were used to calculate the intensity and relative velocity of this line by using a method originally applied to Jupiter[22], and adapted since to Saturn[12,14]. The intensity structure shows three peaks, the outermost two marking the point where the slit cuts through the main auroral oval. The extended flanks of the profile also bulge outwards with additional emission. The velocity structure shows a 'three-tiered' velocity structure[14], with a core region that co-rotates with the planet, flanked by two regions that significantly sub-rotate, which is typical for profiles taken in periods of rarefied solar wind conditions. The errors associated with fitting the data are shown by grey regions surrounding the intensity and velocities, with darker grey where these regions cross. This is the calculated standard deviation of the fitting procedure at each position on the profile.

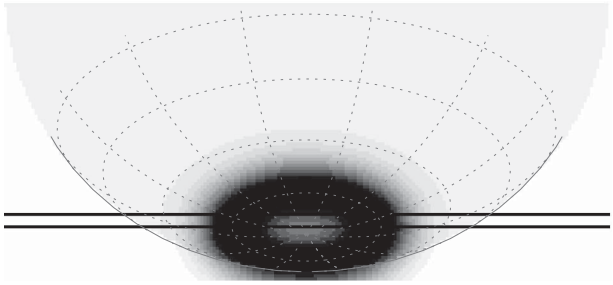


Figure 2 | The location of the spectrometer slit on Saturn's southern aurora. The slit position (bold) was estimated by comparing the auroral data with a modelled main auroral oval. The longitudinal grid is demarcated in steps of 30° and the latitudinal grid in steps of 15°, with the body of the planet (in grey) shown for clarity. The main auroral oval model was based on an axisymmetric model oval that approximates the observed time-averaged ultraviolet auroral oval[15]: ~1° wide, located at 75° latitude. This was convolved to appear as though measured from the ground, broadened by an effective seeing of 1.8 arcsec. The intensity structure, consisting solely of the modelled main auroral oval, thus excludes emission from the breakdown region or diffuse aurora in the modelled intensity.

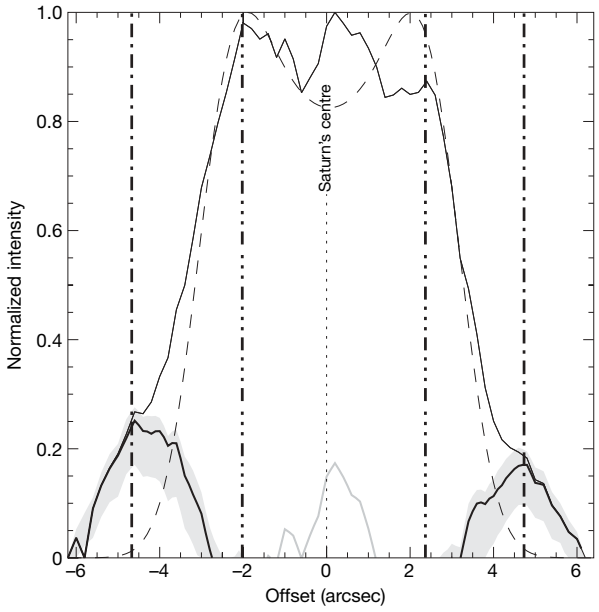


Figure 3 | 'Jovian-like' aurorae on Saturn. The residual intensity on the flanks of the observed oval (bold) and in the polar cap (bold and grey), are found by subtracting the line-of-sight corrected intensity structure, shown in Fig. 1 (thin), from the modelled main auroral oval intensity, shown in Fig. 2 (dashed). The dot–dashed and three-dot–dashed vertical lines are the same as those in Fig. 1. The main auroral oval is modelled as an intensity profile, produced by combining the light entering a slit positioned over the seeing-distorted main auroral oval model. Because the H_3^+ aurora is actually formed within an atmospheric 'shell', where the observed intensities are integrated along a column of atmosphere whose depth varies with position on the planet, we have applied a line-of-sight correction to the spectral data. The residual intensities are calculated by subtracting the modelled main auroral oval from this corrected intensity profile. The errors in the secondary oval intensity are given as a grey region and are the combined standard deviation in fitting the spectra and in positioning the modelled intensity profile. The ultraviolet auroral intensity is relatively stable across the short integration time over which Hubble images are typically taken, as is the infrared aurora when integrated over a period of several hours; however, these do vary over longer timescales. Changing the location of modelled oval equatorward reduces the peak intensity of the secondary oval but never eliminates the secondary aurora completely, and the peak intensity is always located at the breakdown in co-rotation.

wind. However, it is no longer reasonable to consider Saturn's aurorae a 'hybrid' of those of the Earth and Jupiter, but rather that the aurorae of Jupiter and Saturn are variants of the same formation processes.

Received 4 December 2007; accepted 29 April 2008.

1. Dungey, J. W. Interplanetary magnetic field and the auroral zones. *Phys. Rev. Lett.* **6,** 47–48 (1961).
2. Hill, T. W. The Jovian auroral oval. *J. Geophys. Res.* **106,** 8101–8108 (2001).
3. Cowley, S. W. H. & Bunce, E. J. Origin of the main auroral oval in Jupiter's coupled magnetosphere–ionosphere system. *Planet. Space Sci.* **49,** 1067–1088 (2001).
4. Cowley, S. W. H., Bunce, E. J. & O'Rourke, J. M. A simple quantitative model of plasma flows and currents in Saturn's polar ionosphere. *J. Geophys. Res.* **109,** A05212 (2004).
5. Sittler, E. C., Blanc, M. F. & Richardson, J. D. Proposed model for Saturn's auroral response to the solar wind: Centrifugal instability model. *J. Geophys. Res.* **111,** A06208 (2006).
6. Trauger, J. T. *et al.* Saturn's hydrogen aurora: Wide field and planetary camera 2 imaging from the Hubble Space Telescope. *J. Geophys. Res.* **103,** 20237–20244 (1998).
7. Gérard, J.-C. *et al.* Characteristics of Saturn's FUV aurora observed with the Space Telescope Imaging Spectrograph. *J. Geophys. Res.* **109,** A09207 (2004).
8. Clarke, J. T. *et al.* Morphological differences between Saturn's ultraviolet aurorae and those of Earth and Jupiter. *Nature* **433,** 717–719 (2005).
9. Crary, F. J. *et al.* Solar wind dynamic pressure and electric field as the main factors controlling Saturn's aurorae. *Nature* **433,** 720–722 (2005).
10. Stallard, T. *et al.* The H_3^+ latitudinal profile of Saturn. *Astrophys. J.* **521,** L149–L152 (1999).
11. Melin, H., Miller, S., Stallard, T., Trafton, L. M. & Geballe, T. R. Variability in the H_3^+ emission of Saturn: Consequences for ionisation rates and temperature. *Icarus* **186,** 234–241 (2007).
12. Stallard, T., Miller, S., Trafton, L. M., Geballe, T. R. & Joseph, R. D. Ion winds in Saturn's southern auroral/polar region. *Icarus* **167,** 204–211 (2004).
13. Stallard, T. *et al.* Saturn's auroral/polar H_3^+ infrared emission II: A comparison with plasma flow models. *Icarus* **191,** 678–690 (2007).
14. Stallard, T. *et al.* Saturn's auroral/polar H_3^+ infrared emission I: General morphology and ion velocity structure. *Icarus* **189,** 1–13 (2007).
15. Badman, S. V., Cowley, S. W. H., Gérard, J.-C. & Grodent, D. A statistical analysis of the location and width of Saturn's southern auroras. *Ann. Geophys.* **24,** 3533–3545 (2006).
16. Stallard, T., Miller, S., Millward, G. & Joseph, R. D. On the dynamics of the Jovian ionosphere and thermosphere II: The measurement of H_3^+ vibrational temperature, column density, and total emission. *Icarus* **156,** 498–514 (2002).
17. Grodent, D. *et al.* Jupiter's polar auroral emissions. *J. Geophys. Res.* **108,** 1366–1374 (2003).
18. Saur, J. *et al.* Anti-planetward auroral electron beams at Saturn. *Nature* **439,** 699–702 (2006).
19. Grodent, D., Gérard, J.-C., Cowley, S. W. H., Bunce, E. J. & Clarke, J. T. Variable morphology of Saturn's southern ultraviolet aurora. *J. Geophys. Res.* **110,** A07215 (2005).
20. Cowley, S. W. H. & Bunce, E. J. Corotation-driven magnetosphere–ionosphere coupling currents in Saturn's magnetosphere and their relation to the auroras. *Ann. Geophys.* **21,** 1691–1707 (2003).
21. Stallard, T., Miller, S., Cowley, S. W. H. & Bunce, E. J. Jupiter's polar ionospheric flows: Measured intensity and velocity variations poleward of the main auroral oval. *Geophys. Res. Lett.* **30,** 1221–1224 (2003).
22. Stallard, T., Miller, S., Millward, G. & Joseph, R. D. On the dynamics of the Jovian ionosphere and thermosphere I: The measurement of ion winds. *Icarus* **154,** 475–491 (2001).
23. Greene, T. P., Tokunaga, A. T., Toomey, D. W. & Carr, J. S. CSHELL: A high spectral resolution 1–5 micron cryogenic echelle spectrograph for the IRTF. *Proc. SPIE* **1946,** 313–324 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the NASA Infrared Telescope Facility (IRTF) telescope operators for their continued support and expert advice in making these observations possible. The authors are part of the Europlanet European planetary science network, supported by the European Union's Framework 6 programme. This work was supported by the UK Science and Technology Facilities Council, with postdoctoral fellowships for T.S., N.A. and E.J.B., and a senior fellowship for M.D. T.S. is now funded by an RCUK Fellowship. H.M. was supported by a postgraduate studentship from the UK Engineering and Physical Sciences Research Council. S.W.H.C. was supported by a Royal Society Leverhulme Trust Senior Research Fellowship. T.S., H.M. and M.L. are visiting astronomers at the IRTF, which is operated by the University of Hawaii under Cooperative Agreement no. NCC 5-538 with the NASA Science Mission Directorate, Planetary Astronomy Program.

Author Contributions T.S. designed the study, collected and analysed data and wrote the paper. S.M. collected and aided data analysis. H.M. and M.L. aided data analysis. S.W.H.C., E.J.B, N.A. and M.D. provided the magnetospheric context. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.S. (tss@ion.le.ac.uk).

Vol 453|19 June 2008|doi:10.1038/nature07049

LETTERS

Nanoscale holographic interferometry for strain measurements in electronic devices

Martin Hÿtch[1], Florent Houdellier[1], Florian Hüe[1] & Etienne Snoeck[1]

Strained silicon is now an integral feature of the latest generation of transistors and electronic devices[1–3] because of the associated enhancement in carrier mobility[4,5]. Strain is also expected to have an important role in future devices based on nanowires[6] and in optoelectronic components[7]. Different strategies have been used to engineer strain in devices, leading to complex strain distributions in two and three dimensions[8,9]. Developing methods of strain measurement at the nanoscale has therefore been an important objective in recent years but has proved elusive in practice[1,10]: none of the existing techniques combines the necessary spatial resolution, precision and field of view. For example, Raman spectroscopy or X-ray diffraction techniques can map strain at the micrometre scale, whereas transmission electron microscopy allows strain measurement at the nanometre scale but only over small sample areas. Here we present a technique capable of bridging this gap and measuring strain to high precision, with nanometre spatial resolution and for micrometre fields of view[11]. Our method combines the advantages of moiré techniques[12] with the flexibility of off-axis electron holography[13] and is also applicable to relatively thick samples, thus reducing the influence of thin-film relaxation effects.

Transmission electron microscopy (TEM) is the only tool capable of measuring strain at the nanoscale, using techniques based essentially on electron diffraction[14–16]. These techniques have the necessary precision, but they are indirect and rely on detailed comparisons between experimental data and simulation. In addition, they produce isolated point by point measurements and can fail in the highly strained active regions of devices[17,18]. We have previously proposed a combination of high-resolution transmission electron microscopy (HRTEM) and the image processing technique of geometric phase analysis (GPA)[19]. Although this technique is highly accurate at the nanometre scale[20,21], mapping strain in transistors requires large fields of view[22]. To resolve the atomic lattice, high magnifications (typically ×500,000) are required. These images have a resolution of the order of 0.1 nm but the strain information can only be extracted on a scale of 2–3 nm for a precision of 0.1–0.3%, and the field of view is limited (typically 100 nm square). In addition, specimens are necessarily thin, which allows strains to relax with respect to the bulk state[23]. We have therefore searched for a way of measuring strain in thicker samples, at lower magnifications and for larger fields of view without sacrificing precision. The method that we have invented[11] is an optical combination of the moiré technique[12] and off-axis electron holography[13].

The principle of the technique is shown in Fig. 1. A coherent electron beam illuminates the sample in a diffraction condition for a certain set of lattice planes. The sample is composed of a zone of unstrained crystal A of known lattice parameters, adjacent to a zone of strained crystal B, in a similar orientation and diffraction conditions. This geometry resembles most cross-sectional samples of semiconductor thin layers or devices (B) grown epitaxially on a substrate (A). The two diffracted beams can then be interfered with the aid of an electrostatic biprism. Their phase difference can be measured in two dimensions directly from the holographic fringes and will depend on the dynamical elastic scattering and, more importantly, the geometric phase as defined by geometric phase analysis[24]. If the sample is of uniform thickness the former will be a constant phase term, while the latter encodes the strain information through phase gradients[19].

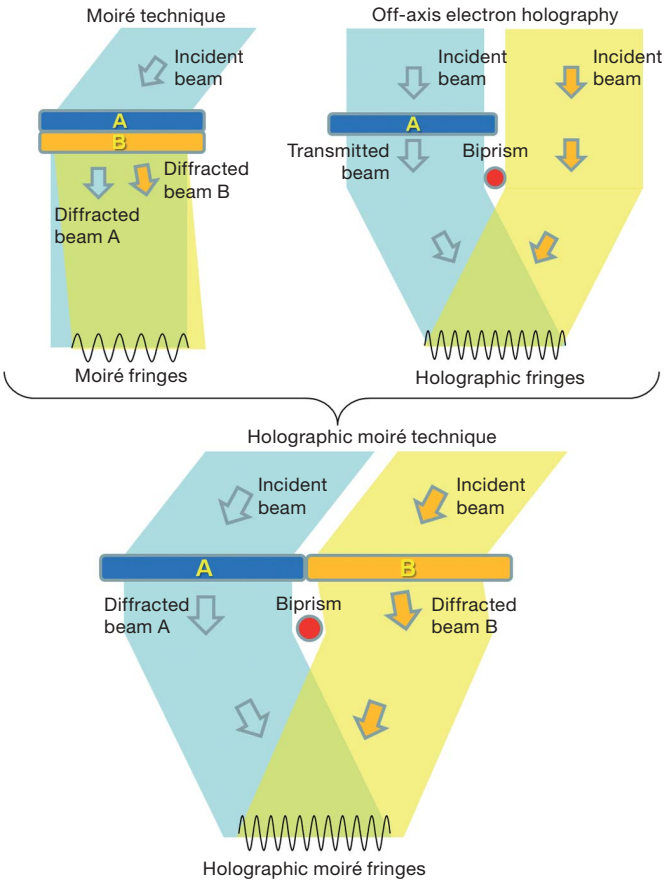


Figure 1 | Principle of the dark-field holographic moiré technique, a combination of the conventional moiré technique and off-axis electron holography. Zone A is an unstrained reference region of crystal, and zone B a region of strained crystal in a similar orientation. The specimen is illuminated with a coherent incident beam, and the diffracted beams (transmitted beams are omitted for clarity) are deviated by the biprism so that they interfere at the screen.

[1]CEMES-CNRS, nMat Group, 29 rue Jeanne Marvig, 31055 Toulouse, France.

The situation is similar to the conventional moiré technique where two samples, one strained and one unstrained, are physically on top of one another (Fig. 1). Although accurate, this technique suffers from two serious drawbacks: first, it is difficult to produce samples with the required geometry, and second, the fringe spacing (which determines the spatial resolution of the measurements) is typically tens of nanometres. The holographic moiré technique solves the problem of sample preparation and nanometric fringes can be obtained. Indeed, the fringe spacing can be tuned to the required spatial resolution by varying the voltage of the biprism. Compared with HRTEM, very large fields of view can be obtained through the use of lower magnifications. We can expect the precision to be similar to HRTEM and possibly higher. The aperture used to isolate the diffracted beams will filter out much unwanted scattered intensity, and stability issues will be reduced because of the wider fringes.

We have tested the technique on a series of dummy transistors (Fig. 2a) with a channel width of 90 nm, using $Si_{80}Ge_{20}$ sources and drains to apply uniaxial compressive stresses to the silicon channel[25]. Figure 2b shows a dark-field hologram obtained by interfering the (220) diffracted beam of the substrate with the active region of the transistor. The phase image (Fig. 2c) is calculated from the hologram and corrected for geometrical distortions[26].

The deformation in the source–drain direction, ε_{xx}, can be calculated directly from the (220) phase because the scattering vector is parallel to the direction in which we wish to measure the deformation. The result is shown in Fig. 3. The field of view, as can be seen, is 0.25 μm wide and over 1 μm long.

To determine the two-dimensional deformation tensor, it is necessary to carry out measurements for at least two lattice planes[19]. We have done this for the (111) and (11$\bar{1}$) diffracted beams and combined the results to obtain the full deformation tensor (Fig. 4). The spatial resolution of the measurements is determined by the fringe spacing and the mask used in the phase reconstruction. Owing to the very high signal-to-noise ratio of the holographic fringes, we were able to obtain the theoretical limit of twice the holographic fringe spacing, that is, 4 nm. To assess precision, we have measured the standard deviation of the strain in a uniform region of substrate and found it to be 0.2%.

To examine the accuracy, we have carried out modelling of the elastic strains using the finite element method[27]. Both the bulk structure (that is, infinite thickness) and a thin TEM sample (5 nm thickness) were simulated. The agreement between the bulk simulations and the experimental results is striking (Fig. 4). For a quantitative comparison, we have extracted profiles from the key deformation maps of lateral source–drain deformation, ε_{xx}, and mean dilatation, δ_{xy} (Fig. 5).

The mean dilatation is sensitive to thin-film relaxation, so it is interesting to note that the experimental data fit the bulk case better than the fully relaxed thin film (Fig. 5a). The sensitivity of the technique is illustrated in the strain profile in the substrate just below the sources and drains (Fig. 5b). Variations of only 0.1% can be detected over distances of nearly 1 μm.

Finally, the most important characteristic for device performance is the value of source–drain compression in the channel[5]. The measurements from the three transistors (Fig. 5c) agree to within 0.1% on average, showing the excellent reproducibility of the technique. The simulated results in the region between 20 and 120 nm from the gate, however, show a systematically lower compression than the experimental curves (by about 0.1%). Rather than attributing this to a thin-film effect, we believe that the modelling of the bulk sample needs to be improved to account better for the experimental data.

The method relies on the preparation of TEM specimens of near uniform thickness, but this is now possible with the development of focused ion beam techniques[28]. Thickness variations are much less important for the strain measurements than, for example, dopant profiling[29]. Geometric phase variations are orders of magnitude larger (12π in Fig. 2c) than the phase changes due to potential changes in doped material (typically π/6)[29]. For the holographic moiré technique the crystal is oriented to a two-beam Bragg diffraction condition and not at a zone axis (as for high-resolution imaging). The

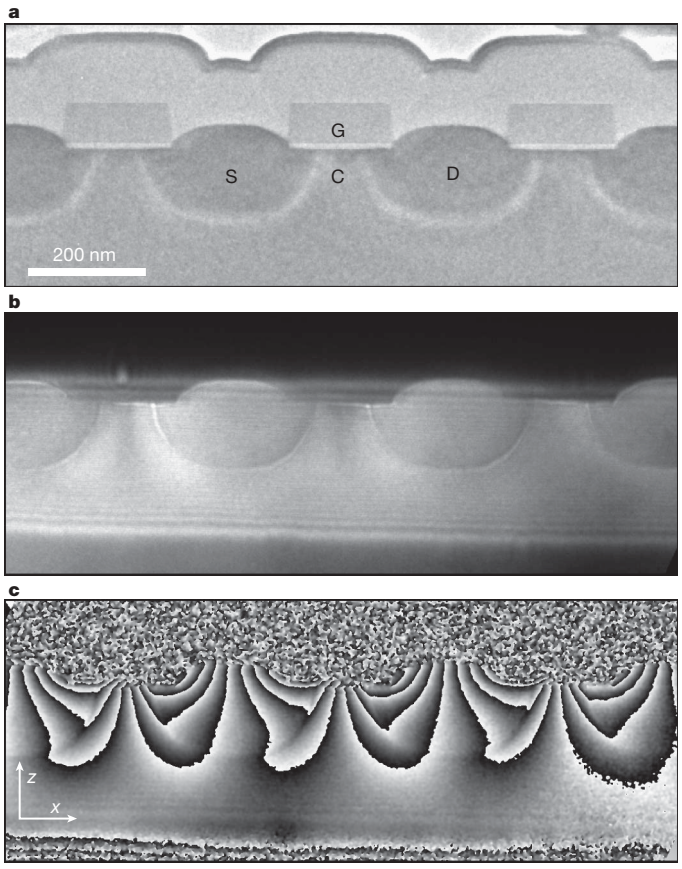


Figure 2 | Dark-field holographic moiré experiment on a strained-silicon transistor array. a, Conventional bright-field image, gate (G), SiGe source (S), SiGe drain (D) and channel (C) indicated. **b**, Dark-field holographic fringes for the (220) diffracted beam. **c**, Phase image calculated from (**b**). Phase is normalized between $-\pi$ (black) and π (white), *x*-axis parallel to [220], *y*-axis parallel to viewing direction [1$\bar{1}$0], and *z*-axis parallel to [002].

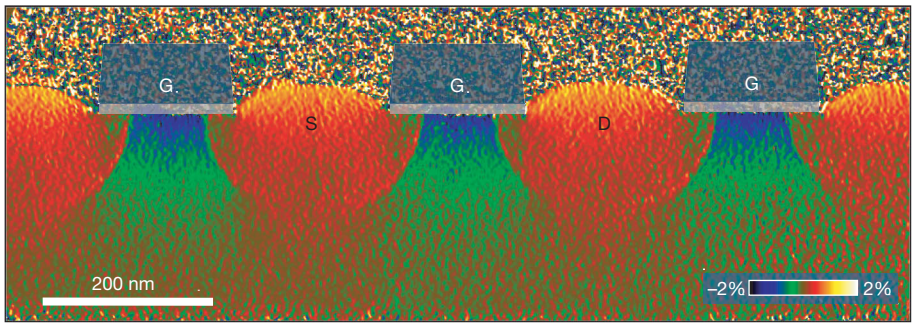


Figure 3 | Measured deformation of strained-silicon transistor array shown in Fig. 2. Deformation in source–drain direction, ε_{xx}, calculated from (220) holographic fringes (Fig. 2c), spatial resolution of 5 nm.

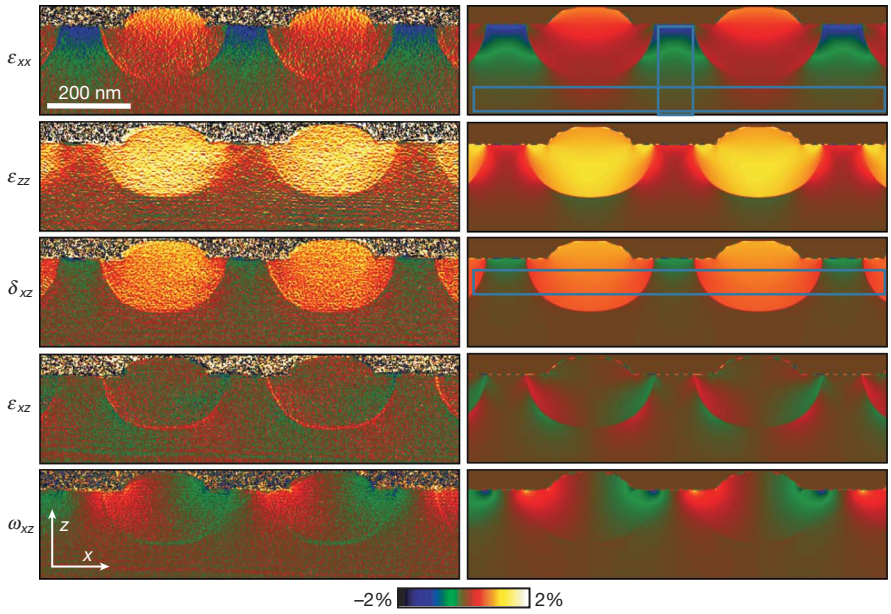
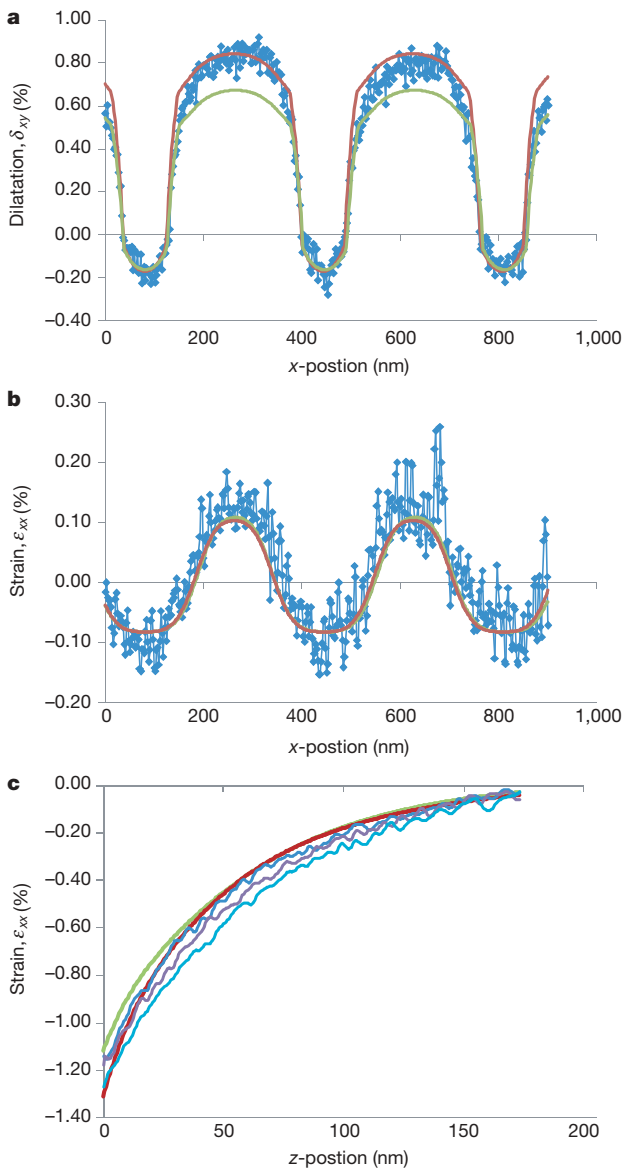


Figure 4 | In-plane deformation maps for transistor array. Left, experimental measurements; right, finite-element modelling of a bulk specimen. The colour scale gives percentage change with respect to the silicon substrate for lattice deformation, ε_{ij}, and degrees for rotation, ω_{xz}. Boxes on simulated results indicate position of Fig. 5 profiles.



phase of the diffracted beam is therefore much less sensitive to thickness variations and to slight changes in specimen orientation.

A particular requirement of the method is that the area of interest be adjacent to a large region of unstrained crystal, but this is the typical geometry of any system composed of a substrate and deposited layers. It is also the case for grain boundaries, where adjacent crystals can be used as references, or precipitates in a matrix. Indeed, samples can be prepared with a suitably oriented crystal next to a region of interest, for example, by the typical method of preparing sandwiched cross-sectional samples.

We believe that the high precision comes from the extreme focus of the method on the experimentally desired quantity—that is, the measurement of the geometric phase. Only one diffracted beam is chosen, and an aperture excludes all extraneous scattering. Indeed, high-order spots can be used to magnify the phase change, which is proportional to the diffraction angle. The large fringe spacing means that measurements are protected from instabilities—it is much easier to image 1-nm fringes than 0.1-nm fringes (as in HRTEM). The only real limitation is the field of view obtainable by holography, and the corresponding fringe spacing and contrast, but these can be optimized using higher brightness electron sources[30], special lens configurations[31] or multiple biprisms[32]. We have shown the extreme sensitivity and wide area possibilities of dark-field electron holography and expect that the technique can be improved significantly with specially designed equipment.

METHODS SUMMARY

We carried out transmission electron microscopy using the SACTEM-Toulouse, a Tecnai F20 ST (FEI) fitted with an imaging aberration corrector (CEOS), rotatable electron biprism (FEI), a 2k CCD (charge-coupled device) camera (Gatan), and imaging filter (Gatan Tridiem). Image analysis was done using a modified version of GPA Phase 2.0 (HREM Research) and DigitalMicrograph

Figure 5 | Deformation profiles of the transistor array. **a**, Mean dilatation profile, δ_{xy}, below the gate across all three transistors, averaged over 60 nm in *z*; **b**, deformation profile, ε_{xx}, in the silicon substrate, averaged over 60 nm in *z*; **c**, vertical strain profiles in *z*, ε_{xx}, taken from gate to substrate for the three transistors and averaged over channel width. Profiles from experimental data (blue), finite-element modelling for bulk sample (red) and thin TEM foil (green). In **a** and **b** the experimental curves show the raw data points every 2 nm (diamonds) joined by a full line, and in **c** the curves are continuous profiles. Positions of profiles and averaging widths are indicated on Fig. 4.

(Gatan) software. Microscope distortions were calibrated to better than 0.05% deformation[26]. We carried out finite-element calculations using the COMSOL Multiphysics (COMSOL) software.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 February; accepted 29 April 2008.

1. *ITRS, International Technology Roadmap for Semiconductors*, 2005 edn Available online at ⟨http://www.itrs.net/reports.html⟩).
2. Ghani, T. *et al.* A 90 nm high volume manufacturing logic technology featuring novel 45 nm gate length strained silicon CMOS transistors. *IEDM Tech. Digest* 978–980 (IEEE International, 2003).
3. Antoniadis, D. A. *et al.* Continuous MOSFET performance increase with device scaling: The role of strain and channel material innovations. *IBM J. Res. Dev.* **50,** 363–376 (2006).
4. Lee, M. L., Fitzgerald, E. A., Bulsara, M. T., Currie, M. T. & Lochtefeld, A. Strained Si, SiGe, and Ge channels for high-mobility metal-oxide-semiconductor field-effect transistors. *J. Appl. Phys.* **97,** 011101 (2005).
5. Thompson, S. E., Sun, G. Y., Choi, Y. S. & Nishida, T. Uniaxial-process-induced strained-Si: Extending the CMOS roadmap. *IEEE Trans. Electron. Dev.* **53,** 1010–1020 (2006).
6. He, R. R. & Yang, P. D. Giant piezoresistance effect in silicon nanowires. *Nature Nanotechnol.* **1,** 42–46 (2006).
7. Jacobsen, R. S. *et al.* Strained silicon as a new electro-optic material. *Nature* **441,** 199–202 (2006).
8. Acosta, A. & Sood, S. Engineering strained silicon: looking back and into the future. *IEEE Potentials* **25,** 31–34 (2006).
9. Parton, E. & Verheyen, P. Strained silicon—the key to sub-45 nm CMOS. *III–Vs Rev.* **19,** 28–31 (2006).
10. Foran, B., Clark, M. H. & Lian, G. Strain measurement by transmission electron microscopy. *Future Fab Intl* **20,** 127–129 (2006).
11. Hÿtch, M. J., Snoeck, E., Houdellier, F. & Hüe, F. Procédé et système de mesure de déformations à l'échelle nanométrique. French Patent Application FR 07 06711.
12. Hirsch, P. B., Howie, A., Nicholson, R., Pashley, D. W. & Whelan, M. J. *Electron Microscopy of Thin Crystals* 2nd edn, ch. 15 (Krieger, Malabar, Florida, 1977).
13. McCartney, M. R. & Smith, D. J. Electron holography: Phase imaging with nanometer resolution. *Annu. Rev. Mater. Res.* **37,** 729–767 (2007).
14. Zhang, P. *et al.* Direct strain measurement in a 65 nm node strained silicon transistor by convergent-beam electron diffraction. *Appl. Phys. Lett.* **89,** 161907 (2006).
15. Usuda, K., Numata, T., Irisawa, T., Hirashita, N. & Takagi, S. Strain characterization in SOI and strained-Si on SGOI MOSFET channel using nano-beam electron diffraction (NBD). *Mater. Sci. Eng. B* **124,** 143–147 (2005).
16. Li, J., Anjum, D., Hull, R., Xia, G. & Hoyt, J. L. Nanoscale stress analysis of strained-Si metal-oxide-semiconductor field-effect transistors by quantitative electron diffraction contrast imaging. *Appl. Phys. Lett.* **87,** 222111 (2005).
17. Clément, L., Pantel, R., Kwakman, L. F. T. & Rouvière, J.-L. Strain measurements by convergent-beam electron diffraction: The importance of stress relaxation in lamella preparations. *Appl. Phys. Lett.* **85,** 651–653 (2004).
18. Houdellier, F., Roucau, C., Clément, L., Rouvière, J.-L. & Casanove, M.-J. Quantitative analysis of HOLZ line splitting in CBED patterns of epitaxially strained layers. *Ultramicroscopy* **106,** 951–959 (2006).
19. Hÿtch, M. J., Snoeck, E. & Kilaas, R. Quantitative measurement of displacement and strain fields from HREM micrographs. *Ultramicroscopy* **74,** 131–146 (1998).
20. Hÿtch, M. J., Putaux, J.-L. & Pénisson, J.-M. Measurement of the displacement field around dislocations to 0.03 Å by electron microscopy. *Nature* **423,** 270–273 (2003).
21. Johnson, C. L. *et al.* Effects of elastic anisotropy on strain distributions in decahedral gold nanoparticles. *Nature Mater.* **7,** 120–124 (2008).
22. Hüe, F., Hÿtch, M. J., Bender, H., Houdellier, F. & Claverie, A. Direct mapping of strain in a strained-silicon transistor by high-resolution electron microscopy. *Phys. Rev. Lett.* **100,** 156602 (2008).
23. Treacy, M. M. J., Gibson, J. M. & Howie, A. On elastic relaxation and long wavelength microstructures in spinodally decomposed $In_xGa_{1-x}As_yP_{1-y}$ epitaxial layers. *Phil. Mag. A* **51,** 389–417 (1985).
24. Hÿtch, M. J. & Plamann, T. Imaging conditions for reliable measurement of displacement and strain from high-resolution electron microscope images. *Ultramicroscopy* **87,** 199–212 (2001).
25. Loo, R. *et al.* A new technique to fabricate ultra-shallow-junctions, combining in situ vapour HCl etching and in situ doped epitaxial SiGe re-growth. *Appl. Surf. Sci.* **224,** 63–67 (2004).
26. Hüe, F. *et al.* Calibration of projector lens distortions. *J. Electron Microsc. (Tokyo)* **54,** 181–190 (2005).
27. Yeo, Y. C. & Sun, J. S. Finite-element study of strain distribution in transistor with silicon-germanium source and drain regions. *Appl. Phys. Lett.* **86,** 023103 (2005).
28. Ishitani, T., Umemura, K., Ohnishi, T., Yaguchi, T. & Kamino, T. Improvements in performance of focused ion beam cross-sectioning: aspects of ion–sample interaction. *J. Electron Microsc.* **53,** 443–449 (2004).
29. Rau, W. D., Schwander, P., Baumann, F. H., Hoppner, W. & Ourmazd, A. Two-dimensional mapping of the electrostatic potential in transistors by electron holography. *Phys. Rev. Lett.* **82,** 2614–2617 (1999).
30. De Jong, N., Allioux, M., Oostveen, J. T., Teo, K. B. K. & Milne, W. I. Optical performance of carbon-nanotube electron sources. *Phys. Rev. Lett.* **94,** 186807 (2005).
31. Wang, Y. Y. *et al.* Off-axis electron holography with a dual-lens imaging system and its usefulness in 2-D potential mapping of semiconductor devices. *Ultramicroscopy* **101,** 63–72 (2004).
32. Harada, K., Akashi, T., Togawa, Y., Matsuda, T. & Tonomura, A. Optical system for double-biprism electron holography. *J. Electron Microsc.* **54,** 19–27 (2005).

Acknowledgements F.H. thanks the CEA-LETI for financial support. This work was partially supported by the European Union through the projects PullNano (Pulling the limits of nanoCMOS electronics, IST-026828) and ESTEEM (Enabling Science and Technology through European Electron Microscopy, IP3: 0260019). We thank P. Mooney for supplying the CCD camera calibration data, P. Verheyen and R. Loo for the device material, and N. Lou and P. Salles for help with FIB preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.H. (hytch@cemes.fr).

METHODS

Specimen preparation. Specimens prepared for electron microscopy are first thinned by the tripod method to obtain 4-μm-thick lamellae. Lamellae are then glued on a half copper grid and placed vertically in the focused ion beam microscope (FIB). We use a Cross Beam XB 1540 FIB (Zeiss): a scanning electron microscope (Gemini Zeiss) combined with a FIB column (Orsay Physics). Optimal imaging resolutions are 2 nm for the scanning electron microscope and 10 nm for FIB. Etching is performed with a 30-keV gallium beam to obtain thin and uniform TEM lamella. First, we use a gallium probe of 1,200 pA to reduce the thickness to below 800 nm (spot size of 1 μm). A gallium probe of 70 pA allows thinning to under 200 nm (spot size of 200 nm) and a 10 pA probe is used for the final polish.

Transmission electron microscopy. We carried out transmission electron microscopy using the SACTEM-Toulouse, a Tecnai F20 ST (FEI) fitted with an imaging aberration corrector (CEOS), rotatable electron biprism (FEI) and imaging filter (Gatan Tridiem). The microscope was operated in the 200-kV pseudo-Lorentz mode[33]. The specimen was oriented to a Bragg condition for the selected diffracted beam. We avoided areas showing strong bend contours. The voltage applied to the electrostatic biprism was 80 V, which produced holographic fringe spacings of 2.0 nm, hologram overlap widths of 250 nm and fringe contrasts of about 8%. Holograms were recorded at a nominal magnification of ×20,000, on a 2k slow-scan CCD camera (Gatan USC 1000) mounted before the imaging filter. Digital sampling densities were 0.566 nm per pixel.

Data analysis. We calculated phase images using a modified version of GPA Phase 2.0 software (HREM Research), a plug-in for the image processing package DigitalMicrograph 3.8+ (Gatan). The holographic fringe periodicity was selected in the Fourier transform of the hologram using a half-cosine mask of radius 0.25 nm^{-1}. No subsequent averaging was carried out in real space. We corrected the phase images for the geometric distortions of the CCD camera using the same method as that prescribed for projector lens distortions[26]. Raw CCD distortion data were supplied by P. Mooney (Gatan). No account was taken of the modulation transfer function of the CCD camera. Deformation maps were calculated by taking numerical derivatives of the phase images using a standard 3 by 3 pixel kernel in real space. Profiles were obtained using the standard tools supplied by DigitalMicrograph.

Modelling. We simulate strains in the p-MOSFET using linear anisotropic elastic theory by resolving partial differential equations with the finite element method[34]. Domains of different chemical composition are distinguished by their elastic coefficients and lattice parameters (determined by applying Vergard's law to the bulk values for silicon and germanium). The geometry of the model is a scrupulous reproduction of the TEM lamella based on the bright-field image (Fig. 2a). We treat epitaxy as a thermal expansion problem[35]. Relaxation of the different domains is principally governed by the elastic tensor in each domain and the boundary conditions. In the x direction, parallel to the transistors, the lamella is considered to be infinite by using periodic boundary conditions. In the z direction of growth, the lower boundary in the substrate is held fixed and the upper surface treated as a free surface. For the bulk simulation, the y direction (electron beam direction) is infinite with periodic boundary conditions. For the thin-film TEM simulation, we consider a 5-nm-thick lamella with two free surfaces.

33. Snoeck, E., Hartel, P., Mueller, H., Haider, M. & Tiemeijer, P. C. Using a CEOS-objective lens corrector as a pseudo Lorentz lens in a Tecnai F20 TEM. *Proc. 16th Intl Microsc. Congress* **2,** 730 (Japanese Society of Microscopy, Sapporo, 2006).
34. Huebner, K. H. H., Dewhirst, D. L., Smith, D. E. & Byrom, T. G. *The Finite Element Method for Engineers* (Wiley, New York, 2001).
35. Christiansen, S., Albrecht, M., Strunk, H. P. & Maier, H. J. Strained state of Ge(Si) islands on Si: Finite element calculations and comparison to convergent beam electron-diffraction measurements. *Appl. Phys. Lett.* **64,** 3617–3619 (1994).

LETTERS

Improved estimates of upper-ocean warming and multi-decadal sea-level rise

Catia M. Domingues[1], John A. Church[1,2], Neil J. White[1,2], Peter J. Gleckler[3], Susan E. Wijffels[1], Paul M. Barker[1] & Jeff R. Dunn[1]

Changes in the climate system's energy budget are predominantly revealed in ocean temperatures[1,2] and the associated thermal expansion contribution to sea-level rise[2]. Climate models, however, do not reproduce the large decadal variability in globally averaged ocean heat content inferred from the sparse observational database[3,4], even when volcanic and other variable climate forcings are included. The sum of the observed contributions has also not adequately explained the overall multi-decadal rise[2]. Here we report improved estimates of near-global ocean heat content and thermal expansion for the upper 300 m and 700 m of the ocean for 1950–2003, using statistical techniques that allow for sparse data coverage[5–7] and applying recent corrections[8] to reduce systematic biases in the most common ocean temperature observations[9]. Our ocean warming and thermal expansion trends for 1961–2003 are about 50 per cent larger than earlier estimates but about 40 per cent smaller for 1993–2003, which is consistent with the recognition that previously estimated rates for the 1990s had a positive bias as a result of instrumental errors[8–10]. On average, the decadal variability of the climate models with volcanic forcing now agrees approximately with the observations, but the modelled multi-decadal trends are smaller than observed. We add our observational estimate of upper-ocean thermal expansion to other contributions to sea-level rise and find that the sum of contributions from 1961 to 2003 is about 1.5 ± 0.4 mm yr^{-1}, in good agreement with our updated estimate of near-global mean sea-level rise (using techniques established in earlier studies[6,7]) of 1.6 ± 0.2 mm yr^{-1}.

To estimate ocean heat content and associated thermosteric sea-level changes from 1950 to 2003 (see Methods), we use temperature data[11] from reversing thermometers (whole period), expendable bathy-thermographs (XBTs; since the late 1960s), modern and more accurate conductivity–temperature–depth (CTD) measurements from research ships (since the 1980s) and Argo floats (mostly from 2001). XBTs, providing more than 50% of the data, measure temperature from free-falling expendable probes at depths estimated from the elapsed time since release at the surface. Significant systematic biases in XBT temperatures[9] are associated primarily with errors in the estimated depth of observations, probably a result of subtle differences in the manufacture of the XBTs[8]. We use a recent time-variable XBT fall-rate correction[8] to minimize these biases. To recover the large-scale patterns from sparse temperature data, we use a reduced-space optimal interpolation technique[5]. This approach helps overcome the low bias in previous estimates of trends in heat content[1,12] and sea level[12,13], particularly in the Southern Hemisphere[3,14], and provides rigorous error estimates (see Methods).

Near-globally averaged anomalies of ocean heat content in the upper 700 m and 100 m (plotted as three-year running means) and sea surface temperature now track each other at multi-decadal timescales (Fig. 1). These global averages all show a slight increase from 1950 to about 1960, a 15-year period to the mid-1970s of zero or slightly negative trend and, after the 1976–1977 climate shift, a steady rise to the end of the record. This pattern is also observed in thermosteric sea level (Fig. 2). Including time-variable error estimates, the linear trend in ocean heat content in the upper 700 m gives a total change of $16 \pm 3 \times 10^{22}$ J from 1961 to 2003 (equivalent to an air–sea heat flux of 0.36 ± 0.06 W m^{-2} over the ocean surface area considered, 3.3×10^{14} m^2; all error estimates quoted are one standard deviation), with about 91% stored in the upper 300 m. For thermosteric

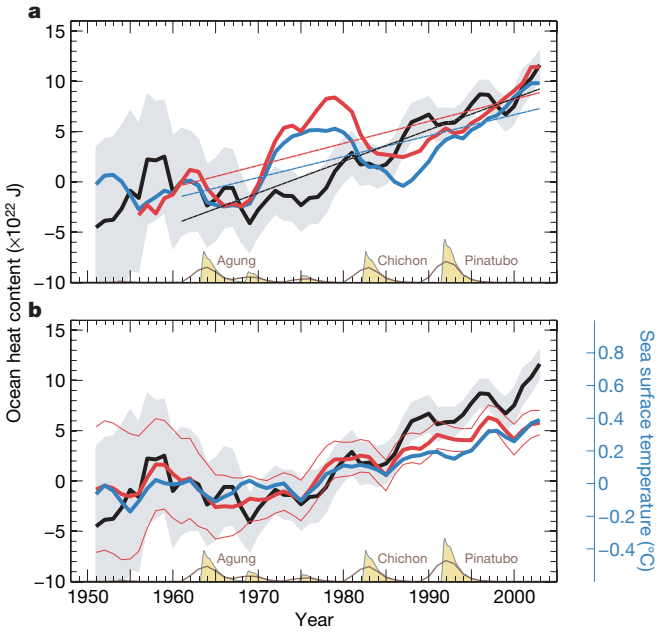


Figure 1 | Estimates of ocean heat content and sea surface temperature. **a**, Comparison of our upper-ocean heat content (black; grey shading indicates an estimate of one standard deviation error) with previous estimates (red[1] and blue[12]) for the upper 700 m. The straight lines are linear fits to the estimates. The global mean stratospheric optical depth[31] (beige, arbitrary scale) at the bottom indicates the timing of major volcanic eruptions. The brown curve is a three-year running average of these values, included for comparison with the smoothed observations. **b**, Comparison of our 700-m (thick black line, as in **a**) and 100-m (thick red line; thin red lines indicate estimates of one standard deviation error) results with sea surface temperature[30] (blue; right-hand scale). All time series were smoothed with a three-year running average and are relative to 1961.

[1]Centre for Australian Weather and Climate Research, CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart, Tasmania 7001, Australia. [2]Antarctic Climate and Ecosystems Cooperative Research Centre, Hobart, Private Bag 80, Tasmania 7001, Australia. [3]Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Mail Code L-103, 7000 East Avenue, Livermore, California 94550, USA.

sea level, the rise is about 22 mm (a trend of 0.52 ± 0.08 mm yr^{-1}, with about 97% in the upper 300 m). These ocean warming and thermal expansion rates are more than 50% larger than previous estimates for the upper 300 m (refs 1, 13) and are about 50% larger for the upper 700 m (refs 1, 12, 13). Since 1976, the equivalent rates have been 0.40 W m^{-2} and 0.59 mm yr^{-1}, and for the period of the modern satellite altimeter record from 1993 to 2003 they have been 0.35 W m^{-2} and 0.79 mm yr^{-1}, less than previous estimates[15], which were biased high by errors in the fall rate of XBTs[8–10]. Exclusion of the XBT data from the reconstructions produces equivalent trends but with larger uncertainties because of the reduced spatial coverage.

Our time series of heat content and thermosteric sea level (Figs 1, 2) now show little indication of the large spurious rise in the early 1970s and the subsequent decrease in the early 1980s (about 6×10^{22} J and about 10 mm), which dominate previous estimates[1,12,13] and were largely the result of instrumental biases[8]. This result is confirmed if we remove the XBT data from the reconstructions. However, there are smaller variations in heat content (less than 3×10^{22} J) and thermosteric sea level (less than 5 mm), roughly consistent in both amplitude and timing with the impact of volcanic eruptions in 1963, 1982 and 1991 (refs 16–18).

We compare our estimates for the upper 300 m (not shown) and 700 m (Fig. 2) with equivalent values from simulations of the twentieth century in the World Climate Research Programme's Coupled Model Intercomparison Project Phase 3 (WCRP CMIP-3) conducted in support of the Intergovernmental Panel on Climate Change Fourth Assessment Report (IPCC AR4). The CMIP-3 simulations examined here are summarized in Supplementary Information. For models that do not include volcanic (stratospheric) aerosols, the changes in simulated ocean heat content and thermosteric sea level have smaller decadal variability than the observations and larger long-term trends (Fig. 2a, b).

For simulations with volcanic forcing (but excluding two models, one that simulates the volcanic forcing by adjusting the solar constant (pale green diamond), and one that responds more strongly to the Agung eruption (orange circle) than the other models[17]), the observed and modelled heat content and thermal expansion time series are similar, with comparable falls in heat content and sea level after volcanic eruptions (Fig. 2c, d). After removal of a linear trend for 1961–1999, the average of the model variances is marginally larger than the variance of the (smoothed) observed time series (which also contains observational uncertainty). The simulated and the observed time series are correlated at zero lag (average correlation coefficient of 0.60). The magnitude of these simulated responses varies because of differences in model physics and estimated volcanic forcings[16–18]. In addition, there are significant differences between ensemble members of the same model.

From 1961 to 1999, the simulations with volcanic forcing have multi-decadal trends in heat storage and thermosteric sea-level rise substantially smaller than those without volcanic forcing. The model trends with volcanic forcing are closer to the observations but are on average about 28% smaller in the upper 300 m and about 10% smaller in the upper 700 m; that is, 73% of the heat storage in the models is in the upper 300 m, in contrast with 93% in the observations.

We combine our estimates of thermosteric sea level with estimates of thermal expansion in the deep ocean and of the increased mass of the ocean in an attempt to balance the sea-level budget (Fig. 3). Although observations and models confirm that recent warming is greatest in the upper ocean, there are widespread observations of warming deeper than 700 m (refs 19–21). The only global observational estimate of thermal expansion in the deep ocean[13] indicates that integrating to 3,000 m gives a 20% increase on the value for the upper 700 m (or 0.07 mm yr^{-1}). This value is probably underestimated because of the use of standard optimal interpolation techniques and the sparse deep observational database, particularly in the Southern Hemisphere[14]. In the ocean reanalysis from the German Consortium for Estimating the Circulation and Climate of the Ocean model, the 1962–2001 ocean thermal expansion was about 0.6 mm yr^{-1} in the upper 700 m, with an additional 50% (about 0.3 mm yr^{-1}) from the ocean below 700 m (ref. 22). For estimating

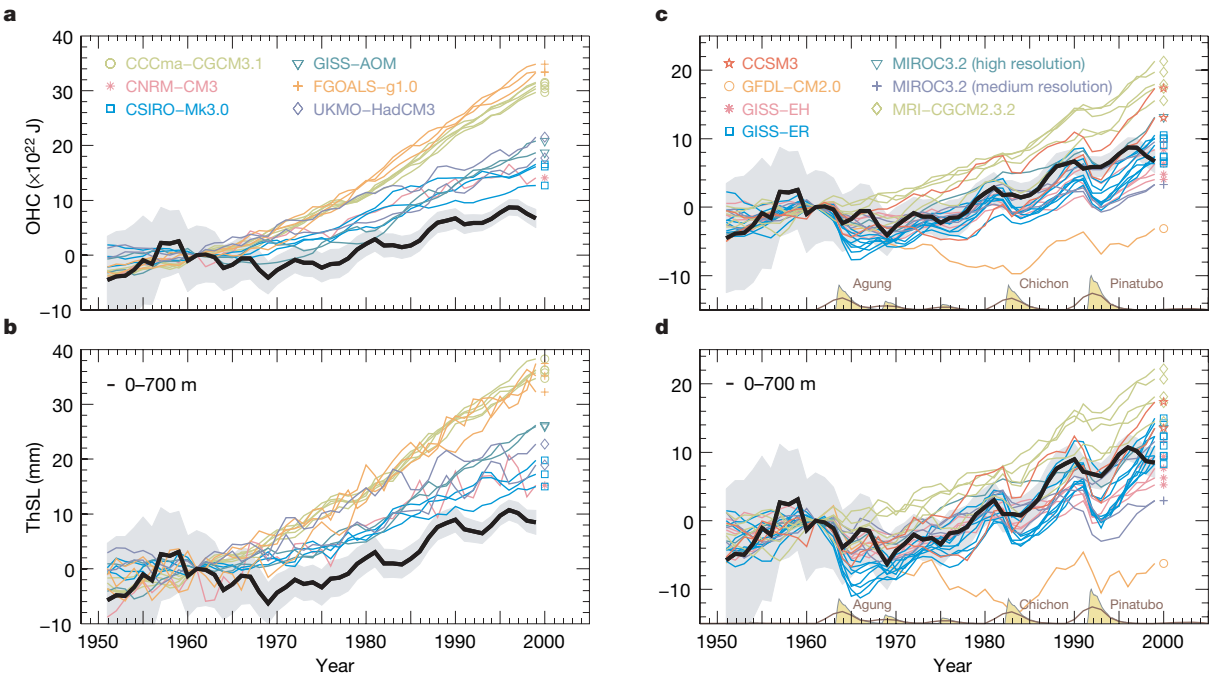


Figure 2 | Comparison of observed and simulated ocean heat content (OHC) and thermosteric sea level (ThSL) estimates for the upper 700 m. **a**, **b**, Models without volcanic forcing. **c**, **d**, Models with volcanic forcing. The observations are smoothed as in Fig. 1 and the model results are yearly averages. All models include greenhouse gas and tropospheric aerosol forcings. See Supplementary Information for more details of the models and the climatic forcings. The stratospheric aerosol loadings[31] of the major volcanic eruptions are shown at the bottom of **c** and **d**. The brown curve is a three-year running average of these values, included for comparison with the smoothed observations. The grey shading indicates estimates of one standard deviation error for the observed time series, and all time series are relative to 1961.

the sea-level budget we use a deep-ocean thermal expansion of 0.2 ± 0.1 mm yr^{-1} (Fig. 3a) but recognize that this value is uncertain. This thermal expansion rate implies additional heat storage of about 8×10^{22} J (0.2 W m^{-2}) in the deep ocean.

For 1961–2003, glaciers and ice caps contribute 0.5 ± 0.2 mm yr^{-1} to global sea-level rise (Fig. 3a), increasing to 0.8 ± 0.2 mm yr^{-1} for 1993–2003 (ref. 23). For 1993–2003, the estimated contributions for the Greenland and Antarctic ice sheets are 0.21 ± 0.07 and 0.21 ± 0.35 mm yr^{-1}, respectively[2]. There is little information to constrain ice sheet contributions for previous decades, but it is thought that the Greenland contribution has increased significantly in recent years[24]. We use a contribution that increases linearly from zero in 1961 to the 1990s value. The Antarctic ice sheet is thought to be still responding to changes since the last glacial maximum[19]. These long timescales suggest that there may have been little change in the Antarctic contribution since 1961.

Hydrological models indicate decadal changes in terrestrial water storage but little long-term trend[25] (Fig. 3a). Terrestrial storage associated with multi-decadal human interference in the water system is poorly determined. The two largest terms, the building of dams (about 0.55 mm yr^{-1} over the past half century[26]) and the mining of groundwater[19], are likely to be of similar size but of opposite sign. For this reason we have not included these terms in Fig. 3.

We update *in situ* estimates of globally averaged sea level by applying established techniques[6,7], but using empirical orthogonal functions determined from a longer set of altimeter data (1993–2006), using a larger number of tide gauges than in previous studies and correcting the tide-gauge data for the impact of atmospheric pressure[27], as well as glacial isostatic adjustment. The globally averaged sea-level trend from this new estimate and one using an independent technique[28] are almost identical from 1961 to 2003, with a trend of 1.6 ± 0.2 mm yr^{-1} (Fig. 3b). The sum of contributions to sea-level rise is 1.5 ± 0.4 mm yr^{-1}, not significantly different from the estimated value. The almost exact agreement in 2003 is fortuitous and the different decadal variability is an indication of the uncertainty in the estimates and the (unknown) variability in the cryospheric and deep-ocean contributions. From 1993 to 2003, the sum of contributions is 2.4 mm yr^{-1}, again almost equal to the estimated trend from tide gauges of 2.3 mm yr^{-1} and still in the upper quartile of the IPCC projections from 1990 (ref. 29). Note that the sea level estimated from satellite altimeter observations follows the *in situ* estimate closely up to 1999 and then begins to diverge, implying a higher rate of rise. It is unclear why the *in situ* and satellite estimates diverge, and careful comparison is urgently needed.

The improved closure of the sea-level budget over multi-decadal periods (Fig. 3b) and the better agreement in the magnitude of observed and simulated decadal variability (Fig. 2c, d) increase confidence in the present results and represent progress since the last two IPCC reports[2,19]. The results indicate an ongoing need for careful quality control of observational data and also for detailed global and regional comparisons of observational estimates with climate models to understand the implications for the detection, attribution and projection of climate change and sea-level rise.

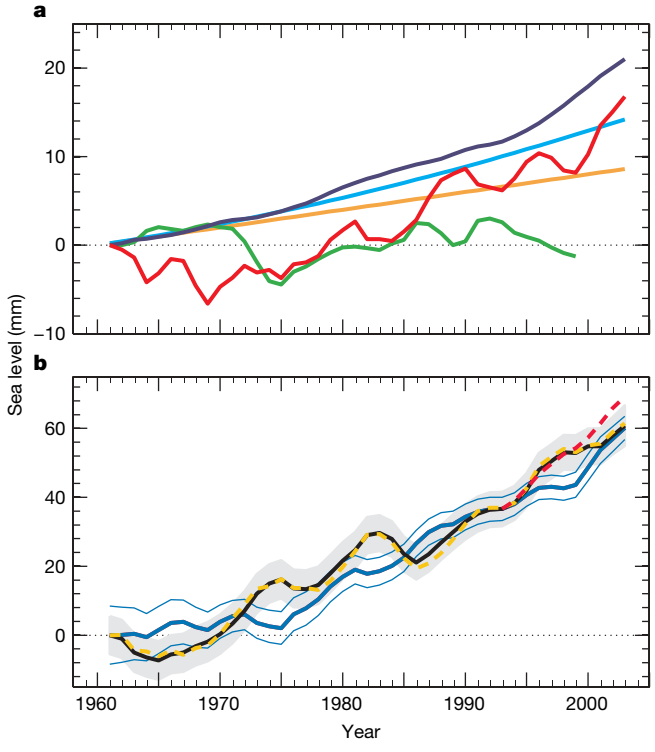


Figure 3 | Total observed sea-level rise and its components. a, The components are thermal expansion in the upper 700 m (red), thermal expansion in the deep ocean (orange), the ice sheets of Antarctica and Greenland (cyan), glaciers and ice caps (dark blue) and terrestrial storage (green). **b**, The estimated sea levels are indicated by the black line (this study), the yellow dotted line[28] and the red dotted line (from satellite altimeter observations). The sum of the contributions is shown by the blue line. Estimates of one standard deviation error for the sea level are indicated by the grey shading. For the sum of components, we include our rigorous estimates of one standard deviation error for upper-ocean thermal expansion; these are shown by the thin blue lines. All time series were smoothed with a three-year running average and are relative to 1961.

METHODS SUMMARY

We reconstructed near-global monthly thermosteric sea level anomalies for 1950–2003 and for different depth levels (100, 200, 300, 400, 500 and 700 m), using a reduced-space optimal interpolation technique[5]. This technique is designed to recover the large-scale robust patterns that can be derived from sparse data and has previously been used to estimate global sea surface temperature[30], atmospheric pressure[27] and sea level[6,7]. The globally averaged time series were computed with equal-area weighting. We converted the thermosteric sea-level fields into changes in ocean heat content using a spatially variable regression. Because of the sparse spatial coverage, particularly in the earlier part of the period, the monthly reconstructed fields contained substantial noise that were reduced by forming three-year running means.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 December 2007; accepted 3 May 2008.

1. Levitus, S., Antonov, J. & Boyer, T. T. Warming of the world ocean, 1955–2003. *Geophys. Res. Lett.* **32,** L02604, doi:10.1029/2004GL021592 (2005).
2. Bindoff, N. L. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 385–432 (Intergovernmental Panel on Climate Change, Cambridge, 2007).
3. Gregory, J. M., Banks, H. T., Stott, P. A., Lowe, J. A. & Palmer, M. D. Simulated and observed decadal variability in ocean heat content. *Geophys. Res. Lett.* **31,** L15312, doi:10.1029/2004GL02058 (2004).
4. AchutaRao, K. M. *et al.* Simulated and observed variability in ocean temperature and heat content. *Proc. Natl Acad. Sci. USA* **204,** 10768–10773 (2007).
5. Kaplan, A., Kushnir, Y. & Cane, M. A. Reduced space optimal interpolation of historical marine sea level pressure. *J. Clim.* **13,** 2987–3002 (2000).
6. Church, J. A., White, N. J., Coleman, R., Lambeck, K. & Mitrovica, J. X. Estimates of the regional distribution of sea-level rise over the 1950 to 2000 period. *J. Clim.* **17,** 2609–2625 (2004).
7. Church, J. A. & White, N. J. A 20th century acceleration in global sea-level rise. *Geophys. Res. Lett.* **33,** L01602, doi:10.1029/2005GL024826 (2006).
8. Wijffels, S. E. *et al.* Changing eXpendable BathyThermograph fall-rates and their impact on estimates of thermosteric sea level rise. *J. Clim.* doi:10.1175/2008JCLI2290.1 (in the press).
9. Gouretski, V. & Koltermann, K. P. How much is the ocean really warming? *Geophys. Res. Lett.* **34,** L01610, doi:10.1029/2006GL027834 (2007).
10. Willis, J., Lyman, J. M., Johnson, G. C. & Gilson, J. Correction to 'Recent cooling of the upper ocean'. *Geophys. Res. Lett.* **34,** L16601, doi:10.1029/2007GL030323 (2007).
11. Ingleby, B. & Huddleston, M. Quality control of ocean temperature and salinity profiles—historical and real time data. *J. Mar. Syst.* **65,** 158–175 (2007).
12. Ishii, M., Kimoto, M., Sakamoto, K. & Iwasaki, S.-I. Steric sea level changes estimated from historical ocean subsurface temperature and salinity analyses. *J. Oceanogr.* **62,** 155–170 (2006).

13. Antonov, J. I., Levitus, S. & Boyer, T. P. Thermosteric sea level rise, 1955–2003. *Geophys. Res. Lett.* **32,** L12602, doi:10.1029/2005GL023112 (2005).
14. Gille, S. T. Decadal-scale temperature trends in the Southern Hemisphere ocean. *J. Clim.* 10.1175/2008JCLI2131.1 (in the press).
15. Willis, J., Roemmich, D. & Cornuelle, B. Interannual variability in upper-ocean heat content, temperature and thermosteric expansion on global scales. *J. Geophys. Res.* **109,** C12037, doi:10.10292003JC002260 (2004).
16. Church, J. A., White, N. J. & Arblaster, J. M. Significant decadal-scale impact of volcanic eruptions on sea level and ocean heat content. *Nature* **438,** 74–77 (2005).
17. Gleckler, P. J. *et al.* Krakatoa lives: The effect of volcanic eruptions on ocean heat content and thermal expansion. *Geophys. Res. Lett.* **33,** L17702, doi:10.1029/2006GL026771 (2006).
18. Delworth, T. L., Ramaswamy, V. & Stenchikov, G. L. The impact of aerosols on simulated ocean temperature, heat content, and sea level in the 20th century. *Geophys. Res. Lett.* **32,** L24709, doi:10.1029/2005GL024457 (2005).
19. Church, J. A. *et al.* in *Climate Change 2001: The Scientific Basis. Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds Houghton, J. T. *et al.*) 639–693 (Cambridge Univ. Press, Cambridge, 2001).
20. Johnson, G. C. & Doney, S. C. Recent western South Atlantic bottom water warming. *Geophys. Res. Lett.* **33,** L14614, doi:10.1029/2006GL026769 (2006).
21. Johnson, G. C., Mecking, S., Sloyan, B. M. & Wijffels, S. E. Recent bottom water warming in the Pacific Ocean. *J. Clim.* **13,** 2987–3002 (2007).
22. Köhl, A., Stammer, D. D. & Cournelle, B. Interannual to decadal changes in the ECCO Global Synthesis. *J. Phys. Oceanogr.* **37,** 313–337 (2007).
23. Dyurgerov, M. B. & Meier, M. F. *Glaciers and the Changing Earth System: A 2004 Snapshot* (Occasional Paper 58, Institute of Arctic and Alpine Research, Univ. of Colorado, 2005).
24. Lemke, P. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) 337–383 (Intergovernmental Panel on Climate Change, Cambridge, 2007).
25. Ngo-Duc, T., Laval, K., Polcher, J., Lombard, A. & Cazenave, A. Effects of land water storage on global mean sea level over the past half century. *Geophys. Res. Lett.* **32,** 9704–9707 (2005).
26. Chao, B. F., Wu, Y. H. & Li, Y. S. Impact of artificial reservoir water impoundment on global sea level. *Science* **320,** 212–214 (2008).
27. Allan, R. & Ansell, T. J. A new globally complete monthly historical mean sea level pressure dataset (HadSLP2): 1850–2004. *J. Clim.* **19,** 5816–5842 (2006).
28. Jevrejeva, S., Grinsted, A., Moore, J. C. & Holgate, S. J. Nonlinear trends and multiyear cycles in sea level records. *J. Geophys. Res.* **111,** C09012, doi:10.1029/2005JC003229 (2006).
29. Rahmstorf, S. *et al.* Recent climate observations compared to projections. *Science* **316,** 709, doi:10.1126/science.1136843 (2006).
30. Rayner, N. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108,** 4407, doi:10.1029/2002JD002670 (2003).
31. Ammann, C. M., Meehl, G. A. & Washington, W. M. A monthly and latitudinally varying volcanic forcing dataset in simulations of the 20th century climate. *Geophys. Res. Lett.* **30,** 16257, doi:10.1029/2003GL016875 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This paper is a contribution to the Commonwealth Scientific Industrial Research Organization (CSIRO) Climate Change Research Program and Wealth from Oceans Flagship and was supported by the Australian Government's Cooperative Research Centres Programme through the Antarctic Climate and Ecosystems Cooperative Research Centre. C.M.D., J.A.C., N.J.W. and S.E.W. were partly funded by the Australian Climate Change Science Program. We acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP-3 multi-model data set. Support for P.J.G. and this data set at the Lawrence Livermore National Laboratory was provided by the Office of Science, US Department of Energy. The Centre for Australian Weather and Climate Research is a partnership between CSIRO and the Australian Bureau of Meteorology.

Author Contributions C.M.D. completed the analysis to determine the changes in ocean heat content and thermosteric sea-level rise and shared responsibility for writing the manuscript. J.A.C. conceived the study, directed the analysis and shared responsibility for writing the manuscript. N.J.W. completed the analysis of the sea-level data and provided the software for the sea-level and thermosteric sea-level reconstructions. P.J.G. analysed the model results. S.E.W. provided the corrections for the XBT data and the climatology, and made valuable comments. P.M.B. provided the pressure corrections to the Argo data. J.R.D. quality-controlled the Argo data. All authors contributed to the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.M.D. (catia.domingues@csiro.au).

METHODS

Ocean temperature data. To estimate ocean heat content for the upper 300 m and 700 m and the associated thermosteric sea level from 1950 to 2003, we used about 2.3 and 1.8 million profiles (shallow and deep, respectively) from the available 6 million ocean temperature profiles in the ENACT/ENSEMBLES version 3 (hereafter EN3) data set[11]. We discarded profiles (about 1.7 million) that had bad quality flags, had coarse vertical resolution, were shallower than 100 m depth or were from higher latitudes than 65° N and 65° S. We carefully selected only the temperature profiles that were clearly identified and measured by XBTs (about 1.2 million and 1.0 million), for which we could apply a correction for the systematic errors[8], bottles (about 1.1 million) and CTDs (about 700,000). We did not include temperature profiles from the remaining instrument types (1.8 million profiles, of which about 1.2 million are from Mechanical Bathy-Thermographs (MBTs)) because of a lack of understanding of their potential systematic biases[8]. To complement the XBTs, bottles and CTDs from the EN3 data set, we used the most recent version of our own quality-controlled Argo profiling floats (about 60,000), including corrections for pressure-sensor drift.

Temperature climatology. We produced a climatology from the observations by using a technique developed previously[32], which includes spatially dependent terms and annual, semi-annual and linear trend terms at each grid point. We believe this is superior to most other available climatologies in which all years are simply averaged together, yielding young median observation dates in the Southern Hemisphere and old median dates in the data-rich areas of the Northern Hemisphere. Attempts to resolve more than a linear trend in time were also considered, but estimates were poorly constrained by the data.

Thermosteric sea level and ocean heat content. We converted temperature profiles into thermosteric sea level and ocean heat content relative to a number of fixed-depth reference levels, assuming climatological salinities from the World Ocean Atlas[33]. We calculated anomalies relative to their monthly mean fields and binned them to a 1 month × 1° × 1° grid for the ice-free ocean equatorward of 65° N and 65° S. Our deepest calculation was performed with respect to 700 m, for comparison with earlier results[1,12,13] and because many XBTs measure to this depth[8]. To take advantage of the greater number of observations in the upper ocean, the 0–700-m estimates are a sum of two depth integrations, 0–300 m and 300–700 m.

Reconstruction details. In our reconstruction we used the sparse but relatively long record of thermosteric sea-level anomalies to determine monthly amplitudes of the leading 30 empirical orthogonal functions (EOFs). The EOFs were used to model variability of the time-varying sea level and were calculated from 14 years (1993–2006) of satellite altimeter data. An additional constant (essentially a spatially uniform field) was included in the reconstruction to represent changes in the global mean[6,7]. Before computing the EOFs, we applied an inverted barometer correction and removed annual and semi-annual signals as well as a globally averaged sea-level trend from the altimeter data.

Error estimates. The reduced-space optimal interpolation formalism[5] provides estimates of errors on the basis of the data distribution and uncertainties in the hydrographic observations (instrumental and geophysical errors) as well as ocean eddy variability determined from satellite altimeter data. The latter two were combined in quadrature. The formal error estimates quoted in the text are for one standard deviation.

Systematic error corrections. We have significantly reduced the systematic biases present in previous analyses by eliminating data sets with unknown errors (for example MBTs), correcting the XBT fall-rate errors and by using the reduced-space optimal interpolation technique. Further refinements in identifying and correcting XBT errors may be possible in the future[8] but it is likely that more than 70% of the earlier XBT biases have been corrected in this analysis. Further corrections are a complex issue that is currently being addressed by an international working group.

Ocean heat content regression. We converted the reconstructed near-global monthly maps of thermosteric sea level into ocean heat content maps by using coefficients obtained from a spatially variable linear regression between estimates of ocean heat content and thermosteric sea level. The regressions are calculated from the temperature profiles in 10° × 10° grid boxes (following the World Meteorological Organization squares). The resultant correlation coefficients are at least 0.99.

32. Alory, G., Wijffels, S. & Meyers, G. M. Observed temperature trends in the Indian Ocean over 1960–1999 and associated mechanisms. *Geophys. Res. Lett.* 34, L02606 10.1029/2006GL028044 (2007).
33. Conkright, M. E. *et al. World Ocean Atlas 2001: Objective Analyses, Data Statistics, and Figures, CD-ROM Documentation* (National Oceanographic Data Center, Silver Spring, MD, 2002).

Vol 453|19 June 2008|doi:10.1038/nature06943

LETTERS

Cytokinin and auxin interaction in root stem-cell specification during early embryogenesis

Bruno Müller[1] & Jen Sheen[1]

Plant stem-cell pools, the source for all organs, are first established during embryogenesis. It has been known for decades that cytokinin and auxin interact to control organ regeneration in cultured tissue[1]. Auxin has a critical role in root stem-cell specification in zygotic embryogenesis[2,3], but the early embryonic function of cytokinin is obscure[4–6]. Here, we introduce a synthetic reporter to visualize universally cytokinin output *in vivo*. Notably, the first embryonic signal is detected in the hypophysis, the founder cell of the root stem-cell system. Its apical daughter cell, the precursor of the quiescent centre, maintains phosphorelay activity, whereas the basal daughter cell represses signalling output. Auxin activity levels, however, exhibit the inverse profile. Furthermore, we show that auxin antagonizes cytokinin output in the basal cell lineage by direct transcriptional activation of *ARABIDOPSIS RESPONSE REGULATOR* genes, *ARR7* and *ARR15*, feedback repressors of cytokinin signalling. Loss of *ARR7* and *ARR15* function or ectopic cytokinin signalling in the basal cell during early embryogenesis results in a defective root stem-cell system. These results provide a molecular model of transient and antagonistic interaction between auxin and cytokinin critical for specifying the first root stem-cell niche.

Cytokinins are adenine-derived signalling molecules that have many essential roles in postembryonic growth and development. However, the role of cytokinin signalling in early embryogenesis remains unclear[4–6]. To visualize cytokinin's signalling output *in vivo*, we aimed to design a synthetic reporter that overcame the limitations of current reporters, typically immediate-early cytokinin target genes. The discrete expression patterns of these markers[7,8] indicate that they integrate unknown secondary input that reflects cytokinin-independent regulation.

Cytokinin signalling is mediated by a multistep two-component circuitry through histidine and aspartate phosphorelay[9]. Nuclear B-type response regulators mediate transcriptional activation in response to phosphorelay signalling activity, whereas A-type response regulators repress signalling in a negative-feedback loop. The DNA-binding domains of diverse B-type response regulator family members are conserved and bind a common DNA-target sequence (A/G) GAT(T/C) *in vitro*[10–12]. This motif is significantly enriched in the *cis*-regulatory region of immediate-early cytokinin target genes[13], suggesting its *in vivo* relevance. To generate a universal cytokinin reporter, we tested and optimized synthetic reporter designs using luciferase (LUC) activity in *Arabidopsis* mesophyll protoplast assays[14,15]. The resulting synthetic reporter, *TCS::LUC* (two-component-output-sensor), harboured the concatemerized B-type Arabidopsis response regulator (ARR)-binding motifs[10–12] and a minimal 35S promoter[14]. Only cytokinins activated *TCS::LUC*, whereas other plant hormones such as auxin, abscisic acid and gibberellic acid, had no effect (Fig. 1a, b). All three known cytokinin receptors contributed to its cytokinin-dependent induction *in vivo*, as cells isolated from double cytokinin receptor mutants were compromised in their ability to induce *TCS::LUC* expression (Fig. 1c). The extent of this reduction correlated with the in planta contribution of the different receptors[6]. B-type ARR family members promoted strong *TCS::LUC* induction in a co-transfection assay[14] (Fig. 1d). Conversely, coexpression of A-type ARR family members inhibited cytokinin-dependent *TCS::LUC* activity (Fig. 1e). *TCS::LUC* displayed concentration-dependent activation by cytokinin from as low as 100 pM up to about 1 μM. Furthermore, *TCS* mediated significantly higher induction compared with the native *ARR6* promoter[14] (Fig. 1f). Addition of a viral translational enhancer (Ω)[16] amplified the response further (Fig. 1f). Taken together, these findings suggest that *TCS::LUC* could specifically report even low levels of phosphorelay output triggered by any of

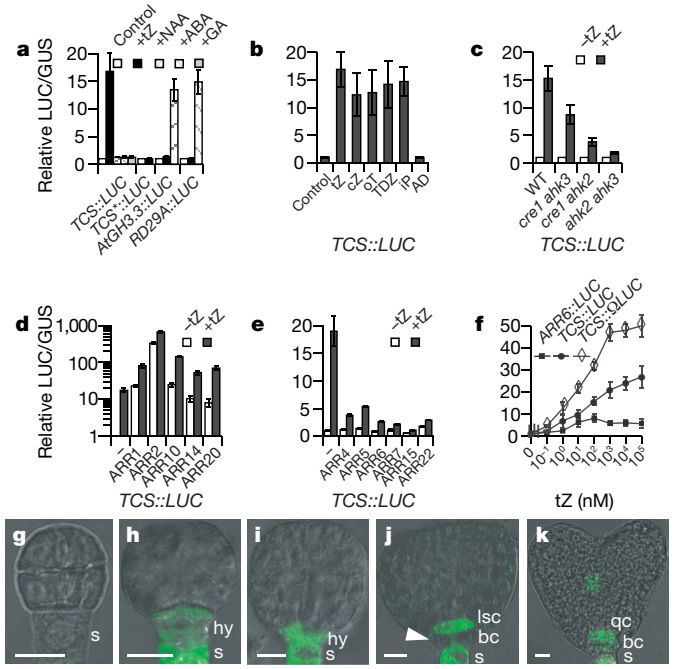


Figure 1 | Sensitive and specific response of *TCS*. **a**, *TCS::LUC* is induced by 100 nM trans-zeatin (tZ), but not 1 μM auxin (NAA), 100 μM abscisic acid (ABA), or 50 μM gibberellic acid (GA). *TCS*::LUC* is a negative control. **b**, *TCS::LUC* is induced by (all at 100 nM): cis-zeatin (cZ), ortho-topolin (oT), thidiazuron (TDZ) and N^6-(Δ^2-isopentenyl)adenine (iP). Adenine (AD) is a negative control. **c**–**e**, *TCS::LUC* induction by trans-zeatin is reduced in double mutant combinations of *ahk2-2*, *ahk3-3* and *cre1-1* (ref. 6) (**c**); stimulated by B-type ARRs (**d**); and reduced by A-type ARRs (**e**). **f**, Dose responses are shown. **g**–**k**, Embryonic *TCS::GFP* activity. Closed arrowhead in (j) points to *TCS::GFP* downregulation in the basal cell lineage. bc, basal cell lineage; hy, hypophysis; lsc, lens-shaped cell; qc, quiescent centre; s, suspensor. Error bars are s.d. ($n = 3$); scale bars represent 10 μm.

[1]Department of Molecular Biology, Massachusetts General Hospital, Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA.

the three endogenous cytokinin receptors and relayed to any response regulator tested.

To determine the expression pattern in planta, we generated transgenic *Arabidopsis* plants carrying the green fluorescent protein (GFP) reporter controlled by the *TCS* synthetic promoter. The activity of *TCS::GFP* in the seedling was consistent with cytokinin actions previously documented, for example, in cotyledons[7] (Supplementary Fig. 1b), shoot meristem[7,17] (Supplementary Fig. 1c), root tip[7,17,18] and root vasculature[19] (Supplementary Fig. 1d), as well as in emerging lateral root base and primordia[20] (Supplementary Fig. 1e, f). Seedlings subjected to a short-term incubation with the cytokinin-synthesis inhibitor lovastatin[21] abolished *TCS::GFP* expression (Supplementary Fig. 2b, e). Notably, *TCS::GFP* expression was restored by the co-administration of a cytokinin together with lovastatin (Supplementary Fig. 2c, f). The analyses validated the physiological response of the novel synthetic cytokinin reporter in intact plants.

To uncover new roles of phosphorelay signalling *in vivo*, we followed *TCS::GFP* expression during early embryogenesis (Fig. 1g–k). As cytokinins have long been implicated in shoot regeneration[1], we were surprised to detect the first distinct signal in the founder of the root stem cells, the hypophysis, at the 16-cell stage (Fig. 1h). By the transition stage, the hypophysis has undergone asymmetrical cell division (compare Fig. 1i with Fig. 1j, and Fig. 2b with Fig. 2c). The resulting large basal daughter cell and its descendants repressed *TCS::GFP* expression, whereas the apical lens-shape cell retained its expression (Figs 1j and 2c). By the heart stage, a second phosphorelay output had appeared near the shoot stem-cell primordium (Fig. 1k).

To identify the signalling components involved in embryonic phosphorelay activity, we determined the transcription levels of all the two-component genes at the transition stage. A subset of candidates for each signalling step was transcribed (Fig. 2a). We were interested in the expression patterns of the A-type ARR genes, as they represent direct targets commonly used as markers of cytokinin signalling. *ARR7* messenger RNA was abundant at the transition stage. *ARR7::GFP* transgenic lines displayed initial GFP activity at the late globular stage, just after the asymmetric division of the hypophysis. Notably, *ARR7::GFP* activity was high in the basal daughter cell but lower in the lens-shape cell and its descendants (Figs 2d and 3c), inverse to *TCS::GFP* levels (Figs 2c and 3a). mRNA *in situ* hybridization confirmed a similar endogenous *ARR7* expression pattern (Fig. 3s). *ARR15*, the sister gene of *ARR7*, is expressed in the comparable domain, albeit at much lower levels (Figs 2a, e and 3d). Therefore, *ARR7* and *ARR15* expression only partially reflected the phosphorelay output (compare Fig. 2d, e with Fig. 2c, and Fig. 3c, d, s with Fig. 3a). These results suggest that additional input might control *ARR7* and *ARR15* expression.

Auxin signalling output visualized by the synthetic reporter *DR5::GFP* (ref. 2) was highest in the hypophysis-derived basal cell (Fig. 2f), similar to *ARR7* and *ARR15* expression. This raised the possibility that auxin could induce transcription of the cytokinin repressors *ARR7* and *ARR15*, which in turn prevented phosphorelay output in the basal cell lineage (Fig. 4m). To test the effect of ectopic auxin signalling, *TCS::GFP, DR5::GFP, ARR7::GFP* and *ARR15::GFP* transgenic embryos (Fig. 3a–d) were treated with the synthetic auxin analogue 2,4-dichlorophenoxyacetic acid (2,4-D; ref. 2) (Fig. 3g–j). Incubation with 2,4-D resulted in expansion of the *DR5::GFP* expression domain (Fig. 3h)[2]. The levels of *ARR7::GFP* and *ARR15::GFP* expression also increased (Fig. 3i, j, t) and their domain widened (Fig. 3i, j), whereas *TCS::GFP* activity was abolished (Fig. 3g). Exogenous cytokinin caused a broad expansion of *TCS::GFP* expression (Fig. 3m), but left *DR5::GFP* expression unaffected (Fig. 3n). Both *ARR7::GFP* and *ARR15::GFP* expression was increased (Fig. 3o, p, t), reflecting their documented status as direct cytokinin target genes[17]. These experiments revealed that besides cytokinin, auxin regulated *ARR7* and *ARR15* transcription; supporting a model where high endogenous auxin activity suppresses cytokinin output through the stimulation of *ARR7* and *ARR15* transcription (Fig. 4m).

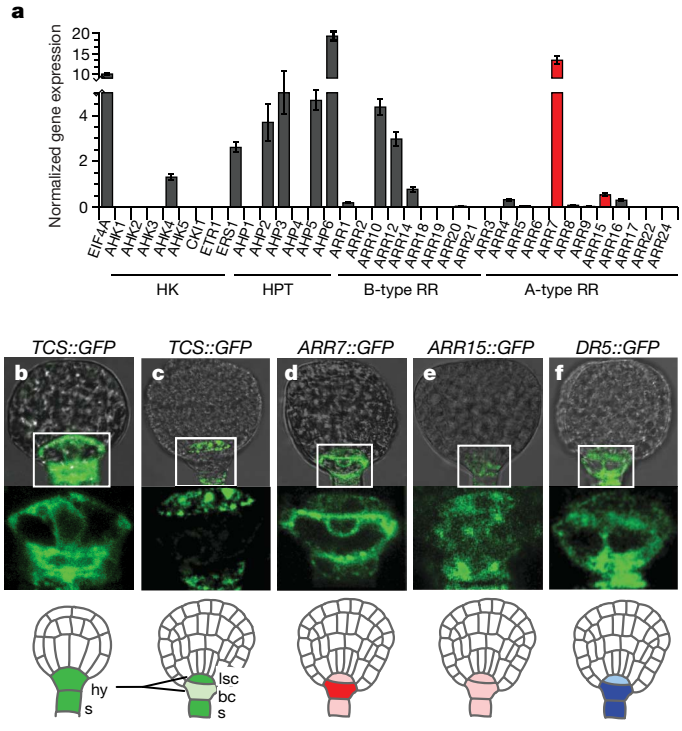


Figure 2 | Inverse correlation between cytokinin and auxin signalling. **a**, Two-component gene transcription by qRT–PCR at transition stage. Error bars represent s.d. ($n = 3$). **b**, *TCS::GFP* in hypophysis (hy) at early globular stage. **c**, Downregulation in basal cell lineage (bc) at late globular stage. **d**, First *ARR7::GFP* expression peaks are in the basal cell lineage. **e**, *ARR15::GFP* expression. **f**, *DR5::GFP* activity is highest in the basal cell lineage. The boxed sections are magnified in the middle panel; a schematic interpretation is shown in the bottom panel. CKI1, cytokinin independent 1; ERS1, ethylene response sensor 1; ETR1, ethylene response 1; HK/AHK, *Arabidopsis* histidine kinases; HPT/AHP, histidine-phosphotransfer proteins; lsc, lens-shaped cell; RR/ARR, response regulators; s, suspensor.

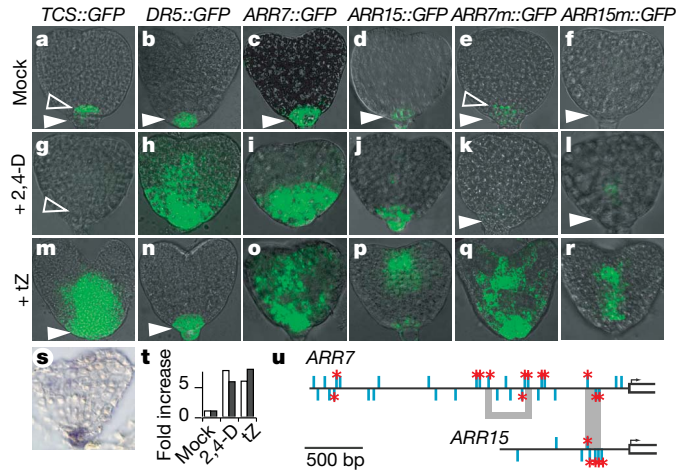


Figure 3 | *ARR7* and *ARR15* mediate auxin control of cytokinin signalling. **a**–**r**, Overnight treatment with water (top row), 50 μM 2,4-D (middle row) or trans-zeatin (tZ; lower row). *ARR7m::GFP* (**e**, **k**, **q**) and *ARR15m::GFP* (**f**, **l**, **r**) are explained in **u**. Filled arrowheads denote basal cell lineage; open arrowheads quiescent centre. **s**, *In situ* hybridization to *ARR7* mRNA. **t**, *ARR7* (open bar) and *ARR15* (filled bar) mRNA induction assayed by qRT–PCR after 4-h treatment of embryos with either 1 μM trans-zeatin or 2,4-D. **u**, Illustration of *ARR7* and *ARR15* promoters. TGTC on sense and antisense strands as blue bars; red asterisks denote point mutations TGTC→TGGC. Grey shading connects conserved regions. See text for details.

To explore the possibility that auxin signalling directly induced transcription of *ARR7* and *ARR15*, we analysed their *cis*-regulatory regions for motifs that might mediate auxin input. Auxin response elements (AuxRE) have been defined as TGTCTC. However, careful *in vitro* analysis demonstrated that only the first four nucleotides are essential for auxin response factor (ARF) binding[22] and we therefore screened the promoters of *ARR7* and *ARR15* for the TGTC motif. We found 32 occurrences in the *ARR7*, and 8 in the *ARR15* upstream region (Fig. 3u). Previously characterized functional *AuxRE* have been categorized as 'simple' (defined by repetitive motifs) or 'composite' (where the motif is flanked by a cofactor-binding site)[23]. These criteria guided us in focusing on putative functional TGTC hits. Point mutations shown to abolish *ARF* binding[22] were introduced specifically in TGTC motifs occurring at least twice in a 30-base-pair (bp) window or flanked by sequence conserved within or between the *ARR7* and *ARR15* promoters (Fig. 3u). The resulting mutated reporters *ARR7m::GFP* and *ARR15m::GFP* (Supplementary Table 2) showed strongly reduced expression in the auxin-signalling domain (Fig. 3e, f, filled arrowhead). Furthermore, ectopic auxin signalling was unable to stimulate their expression (Fig. 3k, l, arrowheads). By contrast, they retained responsiveness to cytokinin (Fig. 3q, r). Notably, uncoupled from auxin input, *ARR7m::GFP* showed an expression pattern similar to that contributed by the cytokinin reporter *TCS::GFP* (compare Fig. 3e with Fig. 3a). Consequently, exogenous auxin application caused repression of *ARR7m::GFP* expression (Fig. 3k, l), probably due to higher endogenous *ARR7* and *ARR15* expression (Fig. 3i, j, t) that prevents cytokinin response (Fig. 3g). These results suggest that auxin signalling directly induces transcription of *ARR7* and *ARR15* through conserved TGTC elements. The sensitivity of the *ARR7* and *ARR15* promoters to auxin seemed to be confined to early embryogenesis, as expression of *ARR7::GFP* and *ARR15::GFP* in the root tip was undetectable by the upturned-U stage (Supplementary Fig. 3), whereas localized auxin signalling persisted.

A question remained as to the function of *ARR7* and *ARR15*, expressed early in embryogenesis under the control of auxin. No embryo defect was observed in the *arr7* or *arr15* single mutants[8] (Fig. 4a–f, Supplementary Fig. 6a–f and data not shown). The *arr7 arr15* double mutants were reported to cause female gametophytic lethality[8], precluding analysis of embryonic function. We therefore generated conditional double loss-of-function *arr7 arr15* embryos by expressing an ethanol-inducible[24] RNA interference

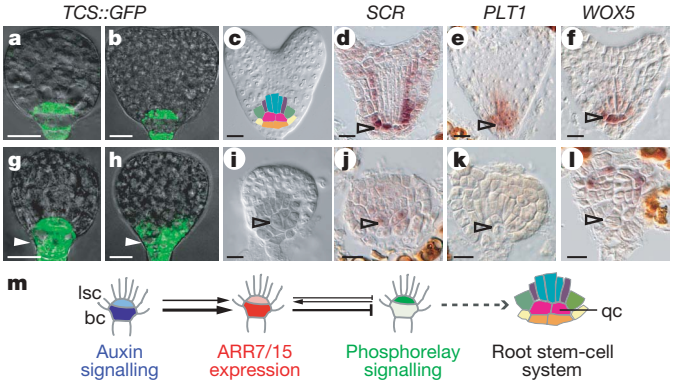


Figure 4 | Function of differential phosphorelay output for root stem-cell establishment. **a–l**, Embryos are *arr15*, *RPS5A::AlcA/AlcR::ARR7(RNAi)*. **a–f**, Control embryos with no ethanol. **g–l**, Embryos after *ARR7(RNAi)* induction in the *arr15* mutant background. Induction was for 10 h (**g**), 36 h (**h**) and 60 h (**i–l**). **d–f, j–l**, *In situ* hybridizations. Artificial colours (**c**) denote stem-cell identity, with quiescent centre (qc) in pink. Cells shaded in grey (**i**) have unclear identity. Filled arrowheads point to basal cell lineage (bc); open arrowheads to qc (**d–f, j–l**) or missing qc (**i–l**). **m**, Model for auxin-dependent phosphorelay downregulation in the basal cell lineage. Scale bars represent 10 μm.

construct against *ARR7* (*ARR7(RNAi)*) (Supplementary Fig. 4) in an *arr15* background with or without the *TCS::GFP* reporter (Fig. 4). Control experiments were performed with single *arr15* mutant embryos carrying uninduced *ARR7(RNAi)* (Fig. 4a–f) and ethanol-induced ARR7(*RNAi*) mutant embryos (Supplementary Fig. 6a–f). Ten hours after *ARR7(RNAi)* transgene induction in *arr15* embryos, ectopic phosphorelay output, revealed by *TCS::GFP* expression, was observed in the basal cell lineage (Fig. 4g). After 36 h, in addition to ectopic cytokinin signalling, cell shapes and number became irregular (Fig. 4h). After 60 h, the morphology of the root stem-cell system was severely distorted (Fig. 4i–l), and the attribution of stem-cell identity based on shape and position was ambiguous in the double mutant (Fig. 4i). Furthermore, the expression of key transcription factors required for root stem-cell specification and function, *SCARECROW* (*SCR*)[25], *PLETHORA 1* (*PLT1*)[26] and *WUSCHEL-RELATED-HOMEOBOX 5* (*WOX5*)[27], was abolished or severely reduced (Fig. 4j–l). Eventually, embryo development arrested (not shown). The single mutant control embryos (Fig. 4a–f and Supplementary Fig. 6a–f) did not show any of these phenotypes. These results suggest that loss of both *ARR7* and *ARR15* causes ectopic cytokinin signalling in the basal cell lineage (Fig. 4g), which interferes with the stereotypical cell division pattern (Fig. 4h) and prevents the establishment of normal embryonic pattern, in particular the root stem-cell system, as judged by morphology (Fig. 4i–l) and expression of key marker genes (Fig. 4j–l).

To determine whether directly activating cytokinin signalling in the basal cell lineage also affects stem-cell development, we used the *DR5* promoter to direct the expression of a constitutively active variant of the B-type *ARR10*, most abundantly expressed in early embryos (Fig. 2a), in auxin-signalling cells. Mutation of the aspartate residue at position 69 to glutamate (D69E) mimics the phosphorylated, active state of *ARR10* (ref. 28). Indeed, early embryonic expression of *ARR10(D69E)* in auxin-maximum cells resulted in a phenotype comparable to loss of *ARR7* and *ARR15* function (Supplementary Fig. 6g). Finally, we addressed the requirement of phosphorelay signalling in early embryo development. It has been reported that mutations in three cytokinin receptors have no obvious effect on embryonic pattern formation[5–7]. Residual activity, or phosphorelay activity independent of known cytokinin receptors, might still occur in these conditions. We chose to interfere dominantly with transcriptional activation executed by B-type ARR proteins and converted the abundant (Fig. 2a) positive regulator *ARR10* into a potent dominant-acting transcriptional repressor by adding an EAR repression domain[29] (Supplementary Fig. 5). Induced ubiquitous expression of *ARR10-EAR* in early globular embryos led to strong pattern defects (Supplementary Fig. 6h). As the lens-shaped cell is prominently marked by *TCS::GFP* expression during early embryogenesis (Fig. 2c), the result suggests that its phosphorelay activity is also important for stem-cell specification. Notably, manipulations of cytokinin signalling initiated later, at embryonic heart stage, had no effect on root stem-cell organization (Supplementary Fig. 7). Thus, differential phosphorelay output seems to be transiently required for successful development of the hypophysis-derived daughter cells into an operational root stem-cell system (Fig. 4m).

By combining a new visualization tool and inducible genetic manipulations, we have uncovered a locally and temporally defined antagonistic interaction between auxin and cytokinin that controls the establishment of the first root stem-cell niche. In the prevailing view, A-type response regulators such as *ARR7* and *ARR15* act in the negative-feedback loop to cytokinin signalling. As a result, A-type response regulator levels are in balance with signalling levels, and their expression domains are centred on the pathway output. By contrast, cytokinin activity will be reduced or eliminated where other signals induce A-type response regulators. Thus, gaining control of feedback regulators represents a simple yet effective mechanism to define the output domain of other pathways, and enables dynamic

and quantitative interactions among signalling pathways to promote the complex plant developmental programmes.

METHODS SUMMARY

Plasmid constructs. *TCS* contains six direct repeats of AAAATCTACAA-AATCTTTTTGGATTTTGTGGATTTTCTAGC (core B-type ARR pentamers[10–12] are underlined); negative control *TCS** has six repeats of AAAATGTA-CAAAATGTTTTTGCATTTTGTGCATTTTCTAGC. *TCS* was cloned in front of a minimal 35S promoter with a TATA box[30], followed by the *LUC*, Ω*LUC* or Ω*GFP*[30] coding regions. The integrated reporter construct with the Ω translational enhancer[16] was designated *TCS::GFP*. The constructs *RPS5A::AlcR/AlcA::ARR7*(*RNAi*) and *DR5rev::AlcR/AlcA::ARR10*(*D69E*) were cloned based on a binary vector received from E. Lam (personal communication). For the *ARR7*(*RNAi*) construct, the first exon and intron of the *ARR7* gene were cloned in sense orientation followed by the first exon in antisense orientation.

Plant material and treatment. Plants were of the Columbia background and grown at 12 h light/23 °C and 12 h dark/20 °C cycle. *In vitro* embryo culture was performed as described[2]. Ethanol (0.5–1.0%), which had no effect on normal embryogenesis, was used to induce transgene expression. The *arr15* (WISCDSLOX334D02) mutant harbours a T-DNA insertion in the first exon of the *ARR15* gene, 73 bp from the translation start (details in Supplementary Fig. 8). The corresponding seed stock CS851593 was obtained from the *Arabidopsis* Biological Resource Centre (ABRC, USA). *Arabidopsis* protoplasts were isolated and transfected as described previously[14,15].

Gene expression analysis. RNA was extracted and amplified from ten pooled embryos, using PicoPure RNA isolation and RiboAMP RNA amplification kits (Arcturus). For expression in wild-type transition-stage embryos, three biological replicates were processed. Quantitative PCR with reverse transcription (qRT–PCR) was performed as described[8] and the amplification of *EIF4A* and *TUB4* served as standards. Primer sequences are provided in Supplementary Table 1.

***In situ* hybridization.** *In situ* hybridization was performed as described previously[8].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 February; accepted 25 March 2008.
Published online 7 May 2008.

1. Skoog, F. & Miller, C. O. Chemical regulation of growth and organ formation in plant tissues cultured *in vitro*. *Symp. Soc. Exp. Biol.* **54,** 118–130 (1957).
2. Friml, J. *et al.* Efflux-dependent auxin gradients establish the apical-basal axis of *Arabidopsis*. *Nature* **426,** 147–153 (2003).
3. Weijers, D. & Jürgens, G. Auxin and embryo axis formation: the ends in sight? *Curr. Opin. Plant Biol.* **8,** 32–37 (2005).
4. Riefler, M., Novak, O., Strnad, M. & Schmülling, T. *Arabidopsis* cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism. *Plant Cell* **18,** 40–54 (2006).
5. Nishimura, C. *et al.* Histidine kinase homologs that act as cytokinin receptors possess overlapping functions in the regulation of shoot and root growth in *Arabidopsis*. *Plant Cell* **16,** 1365–1377 (2004).
6. Higuchi, M. *et al.* In planta functions of the *Arabidopsis* cytokinin receptor family. *Proc. Natl Acad. Sci. USA* **101,** 8821–8826 (2004).
7. To, J. P. *et al.* Type-A *Arabidopsis* response regulators are partially redundant negative regulators of cytokinin signaling. *Plant Cell* **16,** 658–671 (2004).
8. Leibfried, A. *et al.* WUSCHEL controls meristem function by direct regulation of cytokinin-inducible response regulators. *Nature* **438,** 1172–1175 (2005).
9. Müller, B. & Sheen, J. Advances in cytokinin signaling. *Science* **318,** 68–69 (2007).
10. Sakai, H., Aoyama, T. & Oka, A. *Arabidopsis* ARR1 and ARR2 response regulators operate as transcriptional activators. *Plant J.* **24,** 703–711 (2000).
11. Hosoda, K. *et al.* Molecular structure of the GARP family of plant Myb-related DNA binding motifs of the *Arabidopsis* response regulators. *Plant Cell* **14,** 2015–2029 (2002).
12. Imamura, A., Kiba, T., Tajima, Y., Yamashino, T. & Mizuno, T. *In vivo* and *in vitro* characterization of the ARR11 response regulator implicated in the His-to-Asp phosphorelay signal transduction in *Arabidopsis thaliana*. *Plant Cell Physiol.* **44,** 122–131 (2003).
13. Rashotte, A. M., Carson, S. D., To, J. P. & Kieber, J. J. Expression profiling of cytokinin action in *Arabidopsis*. *Plant Physiol.* **132,** 1998–2011 (2003).
14. Hwang, I. & Sheen, J. Two-component circuitry in *Arabidopsis* cytokinin signal transduction. *Nature* **413,** 383–389 (2001).
15. Yoo, S. D., Cho, Y. H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols* **2,** 1565–1572 (2007).
16. Gallie, D. R. The 5′-leader of tobacco mosaic virus promotes translation through enhanced recruitment of eIF4F. *Nucleic Acids Res.* **30,** 3401–3411 (2002).
17. D'Agostino, I. B., Deruere, J. & Kieber, J. J. Characterization of the response of the *Arabidopsis* response regulator gene family to cytokinin. *Plant Physiol.* **124,** 1706–1717 (2000).
18. Aloni, R., Langhans, M., Aloni, E. & Ullrich, C. I. Role of cytokinin in the regulation of root gravitropism. *Planta* **220,** 177–182 (2004).
19. Mähönen, A. P. *et al.* Cytokinin signaling and its inhibitor AHP6 regulate cell fate during vascular development. *Science* **311,** 94–98 (2006).
20. Lohar, D. P. *et al.* Cytokinins play opposite roles in lateral root formation, and nematode and Rhizobial symbioses. *Plant J.* **38,** 203–214 (2004).
21. Orchard, C. B. *et al.* Tobacco BY-2 cells expressing fission yeast cdc25 bypass a G2/M block on the cell cycle. *Plant J.* **44,** 290–299 (2005).
22. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Dimerization and DNA binding of auxin response factors. *Plant J.* **19,** 309–319 (1999).
23. Guilfoyle, T., Hagen, G., Ulmasov, T. & Murfett, J. How does auxin turn on genes? *Plant Physiol.* **118,** 341–347 (1998).
24. Roslan, H. A. *et al.* Characterization of the ethanol-inducible *alc* gene-expression system in *Arabidopsis thaliana*. *Plant J.* **28,** 225–235 (2001).
25. Sabatini, S., Heidstra, R., Wildwater, M. & Scheres, B. SCARECROW is involved in positioning the stem cell niche in the *Arabidopsis* root meristem. *Genes Dev.* **17,** 354–358 (2003).
26. Aida, M. *et al.* The PLETHORA genes mediate patterning of the *Arabidopsis* root stem cell niche. *Cell* **119,** 109–120 (2004).
27. Sarkar, A. K. *et al.* Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* **446,** 811–814 (2007).
28. Hass, C. *et al.* The response regulator 2 mediates ethylene signalling and hormone signal integration in *Arabidopsis*. *EMBO J.* **23,** 3290–3302 (2004).
29. Hiratsu, K., Matsui, K., Koyama, T. & Ohme-Takagi, M. Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. *Plant J.* **34,** 733–739 (2003).
30. Ottenschläger, I. *et al.* Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proc. Natl Acad. Sci. USA* **100,** 2987–2991 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Kakimoto and C. Ueguchi for providing *ahk* mutant seeds; J. Friml for the *DR5::GFP* plasmid and *DR5::GFP* seeds; E. Lam for providing the AlcA/AlcR vector; A. Jazwinska for help with mRNA *in situ* hybridizations; and S. Riku for help with plant growth and protoplast experiments. We also thank C. Ping, S. Howell, Y. Guo, Y. Tan, G. Selvaraj, T. Mizuno, J. Zuo, D. Jackson and J. To for sharing unpublished results. This work was supported by a Fellowship for Prospective Researchers by the Swiss National Science Foundation, a Long Term Fellowship of the International Human Frontier Science Program organization to B.M., and grants from the National Science Foundation and National Institutes of Health to J.S.

Author Contributions B.M. initiated the project, performed the experiments and analysed the data; B.M. and J.S. discussed the results, planned the experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.S. (sheen@molbio.mgh.harvard.edu) or B.M. (mueller@molbio.mgh.harvard.edu).

METHODS

Plasmid constructs. Reporter and effector plasmids used for protoplast transient assays are as previously described[14,15]. The coding regions of *ARR15*, *ARR22* and *ARR20* were obtained from an *Arabidopsis* complementary DNA library by PCR. The *ARR14* gene, the *cis*-regulatory regions of *ARR7* (3 kb), *ARR15* (1.2 kb) and *RPS5A* (1.7 kb)[31] were generated by PCR from genomic DNA. *ARR14*, *ARR15*, *ARR20* and ARR*22* were then cloned into an expression vector as described[14]. Point mutations in the *ARR10* coding region resulting in an aspartate 69 to glutamate mutation, and in the *ARR7* and *ARR15 cis*-regulatory sequences (various TGTC to TGGC mutations) were introduced using the QuikChange Multi Site-Directed Mutagenesis Kit from Stratagene. *ARR7::GFP* and *ARR15::GFP* reporter genes were subcloned into the minibinary vector pCB302 (ref. 32) for plant transformation. To increase the expression levels of reporters, a TMV leader sequence (Ω) stimulating translation[16] was added before the GFP start codon of all GFP constructs. The minimal 35S promoter, Ω and GFP sequence (for *TCS::GFP*), and Ω and GFP sequence (for all other GFP reporters) was amplified by PCR using the *DR5::GFP* (ref. 30) plasmid as a template. The sequences of oligonucleotides used for cloning are provided in Supplementary Table 2. In the ethanol-inducible vector *35S::AlcR/AlcA::*gene of interest (pDM7, gift from E. Lam), the 35S promoter was replaced by the *RPS5A* (ref. 31) or *DR5::Ω* (ref. 30) promoters. After the *AlcA* promoter, *ARR10D69E-GFP*, *ARR10-EAR-GFP* or *ARR7*(*RNAi*) was cloned. The sequences of *ARR7*(*RNAi*), *ARR10*(*D69E*)*-GFP* and *ARR10-EAR-GFP* are provided in Supplementary Table 2. All plasmids were sequenced to ensure that no unwanted mutations were introduced.

Transgenic plants, embryo and seedling analyses. Of fifteen independent transgenic *TCS::GFP* lines screened, three lines with consistent and relatively high expression in embryonic root stem cells were chosen for detailed analysis. Of each stage, at least ten embryos per line were analysed with no variations in expression pattern observed. Typically, transgene silencing in *TCS::GFP* transformed lines was observed beginning in the second generation after transformation, leading to an increased fraction of embryos with reduced or absent GFP activity. Five lines showed very weak expression in embryonic root stem cells whereas seven had no detectable expression in root stem cells. At least six transgenic lines for each GFP construct (*ARR7::GFP*, *ARR15::GFP*, or mutated derivatives) were screened. Two *ARR7::GFP* lines showed relatively weak expression, four lines were intermediate and one line showed stronger expression in embryonic root stem cells. An intermediate line was used for the detailed experiments. Three *ARR7m::GFP* lines had no detectable expression in the root stem cells, two exhibited an expression pattern as reported in this work, and one line exhibited stronger expression. The lines with visible expression were tested for auxin-inducibility as shown in Fig. 3k. None of them showed an increase in expression like the *ARR7::GFP* lines. All of the *ARR15::GFP* lines, six in total, similarly showed very weak GFP expression in root stem cells. All seven of the *ARR15m::GFP* lines had undetectable expression in root stem cells. For *in vitro* embryo culture, a few ovules from each silique were dissected to analyse the stage of the embryos before incubation. The remaining ovules were equally distributed between different treatments and control. All tissue culture plates were sealed with parafilm and kept in the dark overnight for hormone treatment, or up to 60 h for ethanol treatment. To assess the consequences on viability, unopened siliques were incubated up to 10 days in medium containing 0.5× Murashige–Skoog, 0.35% phytagar, 2% sucrose, pH 5.7. Ovules were collected in fixative for mRNA *in situ* hybridizations. Embryos were dissected from ovules and mounted in phosphate buffer to analyse GFP activity, cleared and mounted with chloral hydrate to score phenotypes, or collected in extraction buffer for RNA isolation. Ovules from four independent transgenic lines for *RPS5A::AlcR/AlcA::ARR7*(*RNAi*) in wild-type or *arr15* background were treated with 1% ethanol or incubated without ethanol and assayed in parallel. Most of the *ARR7*(*RNAi*) *arr15* embryos (69%) showed strong defects in root stem cells after 60 h treatment when the ethanol induction started at early globular stage ($n = 71$). About 11% of *ARR7*(*RNAi*) embryos ($n = 55$) and 9% of wild-type ovules ($n = 43$) after 60 h treatment with ethanol showed mild aberrations in the root pole, similar to phenotypes reported previously[33]. The low percentage of mild aberrations was due to embryo culture condition but not ethanol treatment (data not shown). Loss of *PLT1*, *SCR* and *WOX5* was only observed in sections derived from ethanol-induced *ARR7*(*RNAi*) *arr15* embryos. For detailed analysis and crosses to *TCS::GFP*, one of the three *TCS::GFP* lines with high expression in root stem cells was chosen. Ovules from four independent transgenic lines of *RPS5A::AlcR/AlcA::ARR10*(*D69E*), *RPS5A::AlcR/AlcA::ARR10-EAR* and *DR5::AlcR/AlcA::ARR10*(*D69E*) were treated with 0.5% ethanol. Strong phenotypes were observed in 80% of embryos analysed ($n > 40$). Treatment with 1% ethanol increased the severity of the phenotypes in the mutants but not wild-type embryos. To reduce endogenous cytokinin production, lovastatin, a potent inhibitor of the mevalonate pathway[20,34], was prepared as described previously[35] and added to seedlings grown in liquid culture medium (half-strength Murashige–Skoog medium, 1% sucrose, pH 5.7).

***In situ* hybridizations.** The *SCR*, *PLT1* and *WOX5* riboprobes were as described[26,27,36]; the *ARR7* probe comprised the complete *ARR7* translated sequence. Ovules were fixed at 4 °C with 4% paraformaldehyde in PBS for 8 h after vacuum infiltration. The tissue was dehydrated and embedded in paraplast plus. Eight-micrometre sections were placed on SuperFrost-Plus slides. Paraplast was removed by immersion in Histoclear. Sections were rehydrated, incubated for 30 min at 37 °C with 1 μg ml^{-1} proteinase K in TE (50 mM Tris-HCl pH 8, 50 mM EDTA), 10 min in 4% paraformaldehyde in PBS and 10 min in 0.5% acetic anhydride in 0.1 M triethanolamine, pH 8. After dehydration by an ethanol series, slides were air-dried before application of the hybridization solution. Per slide, 50–200 ng labelled riboprobe (probe was hydrolysed in case of *ARR7*) was applied in 80 μl hybridization solution. After incubation in a humidified box at 58 °C overnight (72 h for *ARR7*), slides were washed twice with 2×SSC in 50% formamide for 1 h at 58 °C. Slides were then washed twice in NTE (500 mM NaCl, 10 mM Tris-HCl pH 7.5, 1 mM EDTA) at 37 °C for 5 min each, immersed in preheated (37 °C) buffer 1 (100 mM Tris-HCl pH 7.5, 150 mM NaCl) and then cooled to room temperature. Antibody solution (anti-digoxigenin-alkaline-phosphatase-coupled antibody, diluted 1:2,000 in buffer 1 with 1% blocking reagent) was applied for 2 h. Slides were washed twice for 10 min with 100 mM Tris pH 9.5, 100 mM NaCl. 200 μl of fresh staining solution (10% (w/v) polyvinylalcohol 70–100 kDa, 5 mM MgCl, 0.2 mM 5-bromo-4-chloro-3-indolyl phosphate, 0.2 mM nitroblue tetrazolium salt, 100 mM Tris pH 9.5 and 100 mM NaCl) was added to each slide. Staining occurred over 4 h (72 h for *ARR7* probe) in a humidified box in the dark. Slides were finally washed in water, de- and re-hydrated in ethanol series, and then mounted in 50% glycerol.

Microscopy and imaging. GFP expression was recorded in parallel with transmitted light using a Leica SP2 confocal scanning microscope. Signals were combined in Adobe Photoshop CS3. On the basis of qRT–PCR analysis, the *ARR15* transcript was about 20-fold lower than *ARR7* transcript in transition-stage embryos (Fig. 2a). To visualize the low levels of *ARR15::GFP*, we maximised the sensitivity of the confocal microscope by increasing the signal gain. Embryo sections and cleared whole-mount preparations were recorded with a Leica DFC500 digital camera mounted to a Leica DM5000 microscope.

31. Weijers, D. *et al.* An *Arabidopsis* Minute-like phenotype caused by a semi-dominant mutation in a *RIBOSOMAL PROTEIN S5* gene. *Development* **128**, 4289–4299 (2001).
32. Xiang, C., Han, P., Lutziger, I., Wang, K. & Oliver, D. J. A mini binary vector series for plant transformation. *Plant Mol. Biol.* **40**, 711–717 (1999).
33. Sauer, M. & Friml, J. *In vitro* culture of *Arabidopsis* embryos within their ovules. *Plant J.* **40**, 835–843 (2004).
34. Laureys, F. *et al.* Zeatin is indispensable for the G2-M transition in tobacco BY-2 cells. *FEBS Lett.* **426**, 29–32 (1998).
35. Laule, O. *et al.* Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **100**, 6866–6871 (2003).
36. Di Laurenzio, L. *et al.* The *SCARECROW* gene regulates an asymmetric cell division that is essential for generating the radial organization of the *Arabidopsis* root. *Cell* **86**, 423–433 (1996).

LETTERS

Cortical control of a prosthetic arm for self-feeding

Meel Velliste[1], Sagi Perel[2,3], M. Chance Spalding[2,3], Andrew S. Whitford[2,3] & Andrew B. Schwartz[1–6]

Arm movement is well represented in populations of neurons recorded from the motor cortex[1–7]. Cortical activity patterns have been used in the new field of brain–machine interfaces[8–11] to show how cursors on computer displays can be moved in two- and three-dimensional space[12–22]. Although the ability to move a cursor can be useful in its own right, this technology could be applied to restore arm and hand function for amputees and paralysed persons. However, the use of cortical signals to control a multi-jointed prosthetic device for direct real-time interaction with the physical environment ('embodiment') has not been demonstrated. Here we describe a system that permits embodied prosthetic control; we show how monkeys (*Macaca mulatta*) use their motor cortical activity to control a mechanized arm replica in a self-feeding task. In addition to the three dimensions of movement, the subjects' cortical signals also proportionally controlled a gripper on the end of the arm. Owing to the physical interaction between the monkey, the robotic arm and objects in the workspace, this new task presented a higher level of difficulty than previous virtual (cursor-control) experiments. Apart from an example of simple one-dimensional control[23], previous experiments have lacked physical interaction even in cases where a robotic arm[16,19,24] or hand[20] was included in the control loop, because the subjects did not use it to interact with physical objects—an interaction that cannot be fully simulated. This demonstration of multi-degree-of-freedom embodied prosthetic control paves the way towards the development of dexterous prosthetic devices that could ultimately achieve arm and hand function at a near-natural level.

Two monkeys were implanted with intracortical microelectrode arrays in their primary motor cortices. Each monkey used the signals to control a robotic arm to feed itself. The robotic arms used in these experiments had five degrees of freedom: three at the shoulder, one at the elbow and one at the hand. Like a human arm, they permitted shoulder flexion/extension, shoulder abduction/adduction, internal/external rotation of the shoulder and flexion/extension of the elbow. The hand consisted of a motorized gripper with the movement of its two 'fingers' linked, providing proportional control of the distance between them. Monkeys were first trained to operate the arm using a joystick (Supplementary Methods). Their own arms were then restrained and the prosthetic arm was controlled with populations of single- and multi-unit spiking activity from the motor cortex. The neural activity was differentially modulated when food was presented at different target locations in front of the monkey. Based on previous work[24], we used this modulation to represent velocity of the prosthetic arm's endpoint (a point between the fingertips of the hand/gripper) as an expression of the intention to move[2,3]. The recorded signal was also used by the subject to open and close the gripper as it grasped and moved the food to the mouth. The endpoint velocity and gripper command were extracted from the instantaneous firing rates of simultaneously recorded units using a real-time extraction algorithm.

Many algorithms of varying complexity have been developed in open-loop[7,25–27] or closed-loop experiments[12–24], but here we show that a simple algorithm functioned well in this application. The population vector algorithm[28] (PVA) used here was similar to algorithms used in some cursor-control experiments[15,21]. It relies on the directional tuning of each unit, characterized by a single preferred direction in which the unit fires maximally. The real-time population vector is essentially a vector sum of the preferred directions of the units in the recorded population, weighted by the instantaneous firing rates of the units, and was taken here to represent four dimensions—velocity of the endpoint in an arbitrary extrinsic three-dimensional cartesian coordinate frame, and aperture velocity between gripper fingers (fourth dimension). The endpoint velocity was integrated to obtain endpoint position, and converted to a joint-angular command position, for each of the robot's four degrees of freedom, using inverse kinematics. Degree-of-freedom redundancy was solved by constraining elbow elevation in a way that resulted in natural-looking movements (Supplementary Methods). As the monkey's cortical command signal was decoded in small time-increments (30 ms), the control was effectively continuous and the animal was able to continuously change the speed and direction of arm movement and gripper aperture. Details of the control algorithm are in Supplementary Methods.

To demonstrate fully embodied control (Fig. 1), monkeys learned a continuous self-feeding task involving real-time physical interaction between the arm, a food target, a presentation device (designed to record the target's three-dimensional location) and their mouth. Unlike short control windows used in previous studies, each monkey controlled the arm and gripper continuously during an entire session (not only during reaching and retrieval movements but also during loading/unloading and between trials). The task was challenging owing to the positional accuracy required (about 5–10 mm from the target centre position at the time of gripper closing). The required accuracy for retrieval was much lower because the monkey could move its head to meet the gripper. Supplementary Video 1 shows monkey A performing seven consecutive successful trials of continuous self-feeding. It can be seen from the video that the monkey was still chewing on the previous piece of food while reaching for the next one. It can also be seen that the monkey was able to move its head and eyes naturally without affecting control of the prosthetic arm. Example signals from the last four trials of the video show the correspondence between the spike signals of the 116 units used for control during that session and the resulting arm and gripper movement (Fig. 2).

Monkey A performed 2 days of the continuous self-feeding task with a combined success rate of 61% (67 successes out of 101 attempted trials on the first day, and 115 out of 197 on the second day). To put this success rate in perspective, a task of comparable difficulty to a previous virtual cursor control study from our group[15] would be to simply move the prosthetic arm's endpoint near the target

[1]Department of Neurobiology, School of Medicine, E1440 BST, Lothrop Street, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. [2]Department of Bioengineering, 749 Benedum Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA. [3]Center for the Neural Basis of Cognition, University of Pittsburgh and Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [4]Department of Physical Medicine and Rehabilitation, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. [5]McGowan Institute for Regenerative Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15219, USA. [6]Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.

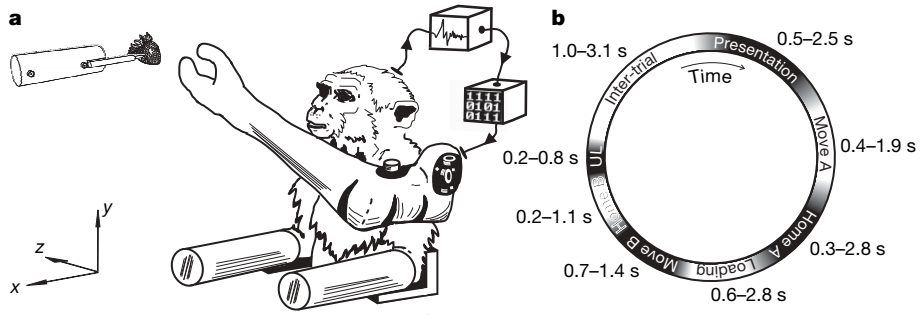


Figure 1 | Behavioural paradigm. a, Embodied control setup. Each monkey had its arms restrained (inserted up to the elbow in horizontal tubes, shown at bottom of image), and a prosthetic arm positioned next to its shoulder. Spiking activity was processed (boxes at top right) and used to control the three-dimensional arm velocity and the gripper aperture velocity in real time. Food targets were presented (top left) at arbitrary positions. **b**, Timeline of trial periods during the continuous self-feeding task. Each trial started with presentation of a food piece, and a successful trial ended with the monkey unloading (UL) the food from the gripper into its mouth (see Methods). Owing to the continuous nature of the task, there were no clear boundaries between the task periods.

(that is, complete the Move A period only, without being required to home in, load, retrieve and unload). (The Move A period is defined in Methods, and shown within the timeline in Fig. 1b.) Monkeys in that previous study had a success rate of 80%, whereas our monkey A successfully completed the Move A period in 98% of attempted trials (Supplementary Table 3). Distance of the targets in this task (184 ± 31 mm, mean ± s.d.) was also greater than that in the previous study. Monkey P performed a version of the continuous self-feeding task (Supplementary Video 2) with an average success rate of 78% (1,064 trials over 13 days), typically using just 15–25 cortical units for control. Monkey P's success rate was higher than monkey A's because the task was easier (see Supplementary Methods).

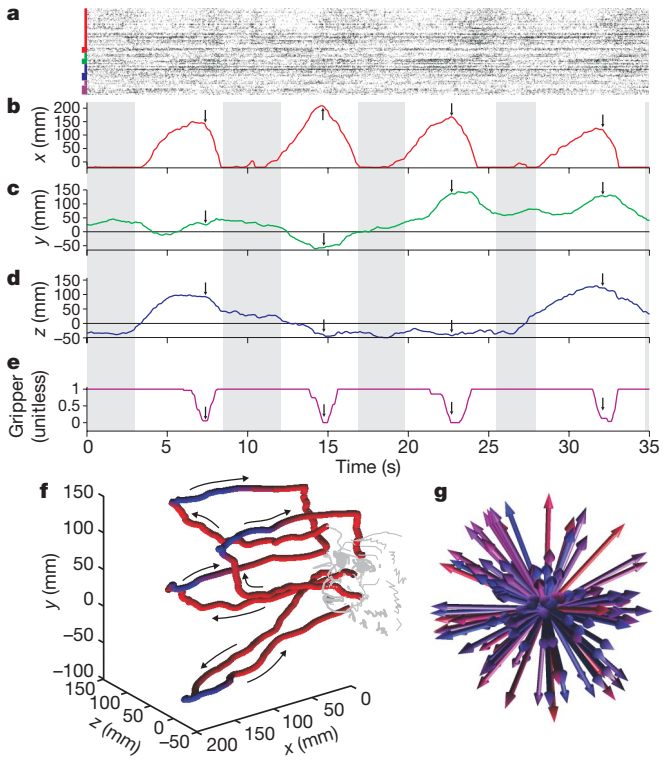


Figure 2 | Unfiltered kinematic and spike data. a, Spike rasters of 116 units used for control. Rows represent spike occurrences for each unit, grouped by major tuning component (red, *x*; green, *y*; blue, *z*; purple, gripper). Groups are further sorted by negative major tuning component (thin bar) versus positive (thick bar). **b**–**d**, The *x*, *y*, and *z* components, respectively, of robot endpoint position. Grey background indicates inter-trial intervals. Arrows indicate gripper closing at target. **e**, Gripper command aperture (0, closed; 1, open). **f**, Spatial trajectories for the same four trials. Colour indicates gripper aperture (blue, closed; purple, half-closed; red, open). Arrows indicate movement direction. **g**, Distribution of the four-dimensional preferred directions of the 116 units used. Arrow direction indicates *x*, *y*, *z* components, colour indicates gripper component (blue, negative; purple, zero; red, positive).

The fact that the gripper opens and closes fully each time (Fig. 2e) indicates good performance, because full opening is advantageous on approach to target and full closing is required for loading. The fact that the task requirements allow the monkey to drive the gripper aperture to both limits makes this fourth dimension easier to control than the *x*, *y* and *z* dimensions. However, the monkey is capable of partially opening or closing the gripper, as shown by data from an earlier training session (Supplementary Fig. 12).

Figure 2f reveals a surprising point: after gripping the food and pulling it off the presentation device, the monkey gradually opened the gripper on the way back to the mouth (Move B) and the gripper was typically fully open before it reached the mouth. One might expect the food to have dropped when the gripper was opened, but this was not always the case because marshmallows, and even grape halves to some extent, tended to stick to the gripper fingers. In an earlier training session, the monkey kept the gripper closed all the way back to the mouth (Supplementary Fig. 13). Over the course of training, the monkey must have learned that keeping the gripper closed was unnecessary, illustrating the importance of working within a physical environment.

We assume that an arm that moves naturally with a bell-shaped speed profile[29,30] will be easier to control than one that moves in an unfamiliar way. Monkey A's individual-trial profiles (Fig. 3a) show a large bell-shaped peak for retrieval movements. Reaching movements consist of multiple smaller bell-shaped peaks indicative of corrective movements. The speed profiles shared qualitative characteristics with natural movements, but the duration of prosthetic movements (3–5 s for monkey A, including reaching, loading and retrieval) is not yet down to the same level as natural movements (1–2 s). The corrective movements and long movement duration are consistent with extensive use of visual feedback in this task.

The animal controlled the exact path of the arm to achieve the correct approach direction to position the gripper in the precise location needed to grasp the food. This was demonstrated by the curved path taken to avoid knocking the food piece off the presentation device (Fig. 3b and Supplementary Video 3). It is also important that there be no apparent control delay—that is, lag between the desire to move and the movement of the prosthetic. The delay between spike signals and movement of the robotic arm was approximately 150 ms (Supplementary Methods). This is not very different from the control delay of a natural arm[6]. An example of lag-free control can be seen in Supplementary Video 2, where the food

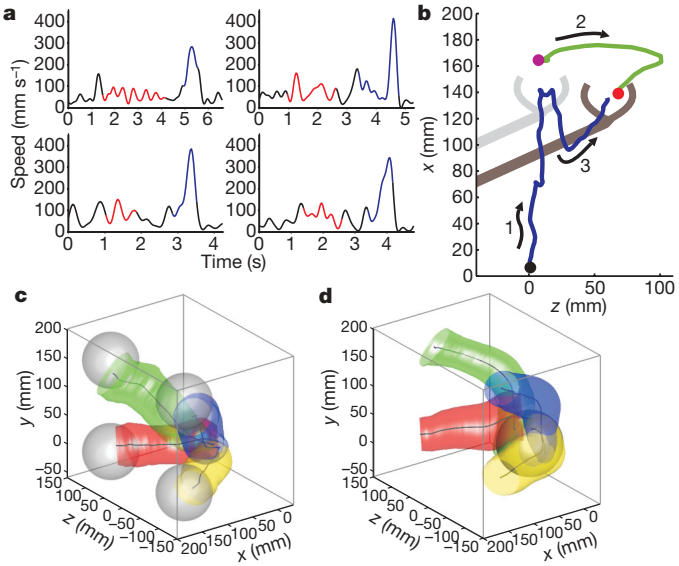


Figure 3 | Movement quality. **a**, Speed profiles from four trials. Time zero marks the beginning of forward arm movement. Reaching (red) begins when the target is in position and ends when the gripper touches the target or minimal distance between target and endpoint is achieved (whichever comes first). Retrieval (from food off the presentation device to mouth contact) is blue, and the graph ends with food in the monkey's mouth (obtained from video record). **b**, Target tracking. Endpoint trajectory (blue, arrow 1) from an initial position (black dot) towards an initial target (purple dot). When the gripper was about to arrive (light grey sketch) at the initial target, the target was shifted (green trajectory, arrow 2) to a new position (red dot). The monkey then moved the arm in a curved path (arrow 3) to avoid knocking the food off the presentation device, positioning the gripper (dark grey sketch) to grasp the food. This trial is also shown in Supplementary Video 3. **c**, **d**, Endpoint trajectory variability (monkey A) for reaching (**c**, Move A period) and retrieval (**d**, Move B). Semi-transparent coloured regions represent trajectory standard deviation (over all sessions) around average trajectories (grey lines) to each target. Grey spheres (radius 46 mm, averaged over all sessions) represent regions where training assistance was applied.

dropped out of the gripper unexpectedly during a retrieval movement and the animal immediately stopped moving the arm.

Some displays of embodiment would never be seen in a virtual environment. For example, the monkey moved the arm to lick the gripper fingers while ignoring a presented food target (Supplementary Video 4), and sometimes used the gripper fingers to give a second push to the food when unloading (Supplementary Video 5). These behaviours were not task requirements, but emerged as new capabilities were learned, demonstrating how the monkey used the robot arm as a surrogate for its own.

The monkeys' arms were restrained in these experiments to prevent them from grabbing the food directly with their own hands. The restraints did not prevent them from making small wrist and hand movements. In particular, monkey A can be seen making characteristic movements with its right hand (Supplementary Video 1): extending the wrist and fingers while closing the prosthetic gripper, then rotating its wrist and flexing the fingers while retrieving the food with the prosthetic arm. It could be argued that these movements might facilitate prosthetic control. However, there are several reasons we find this unlikely. First, the electrode array was implanted in the right hemisphere (the same side as the monkey's own moving hand), while predominant motor cortical output projects to the opposite side of the body. Second, the monkey's hand movement was only loosely coupled to prosthetic control. For example, the temporal correspondence between wrist extension and gripper closing varied between zero and almost a full second (Supplementary Table 4). Third, movement is not required for brain-controlled tasks, as monkeys in other studies made no movement with their arms[14,15,17,22] and paralysed humans have well modulated motor cortical activity capable of driving prosthetic devices[12,13,20]. The arm and hand movements seen here may be vestigial, remnants of the joystick task carried out during initial training.

As an intermediate training step towards continuous self-feeding, after the monkeys learned to operate the device with a joystick, they performed an assisted brain-controlled task where the monkey's control was mixed with automated control. The types and amounts of assistance were configurable in each task period. For example, during the Home A and Loading periods (defined in Methods), the training program partially guided the endpoint towards the target by adding a vector pointing towards the target to the endpoint velocity. Gripper opening was partially aided during Move A and Home A by adding a positive value to aperture velocity, and closing was aided during Loading by adding a negative value. Monkey P also used another type of assistance, where the amount of deviation from a straight line towards the target was limited by a gain factor. The relative proportion of all types of automated assistance in the overall control signal was reduced over several weeks until both the arm endpoint movement and gripper were controlled purely by the monkey's cortical command. Full details of assisted control are in Supplementary Methods. Targets during the training period were presented at four discrete locations. This allowed a measure of trajectory consistency to be computed over repeated trials (Fig. 3c and d). Like natural arm movements, the reaching and retrieval movements of the prosthetic arm show some variability, but are generally consistent between trials.

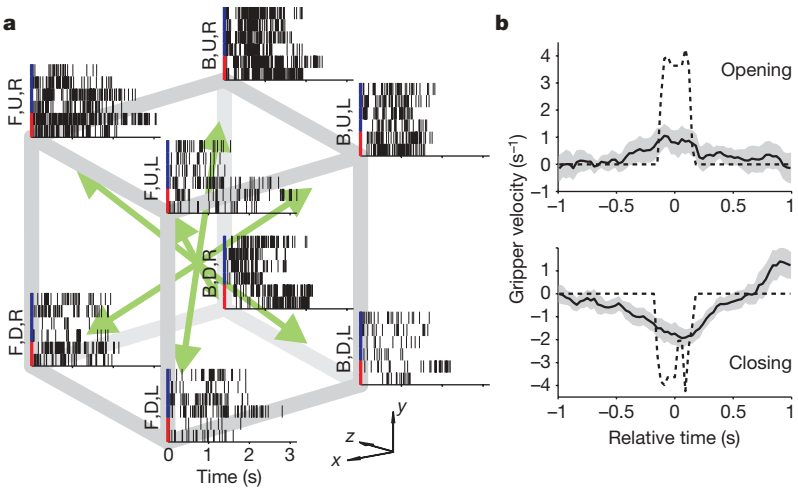


Figure 4 | Unit modulation. **a**, Spike rasters of a single unit during six movements in each of eight directions. This unit (with $\{x,y,z\}$ components of its preferred direction, PD = $\{-0.52,0.21,0.47\}$) fired maximally in the backward-up-right direction (B,U,R) while retrieving from the lower left target, and fired least in the forward-down-left direction (F,D,L) while reaching to the same target. The modulation was consistent during (blue side bars) and after calibration (red side bars). **b**, Gripper modulation. Aperture command velocity (dotted line) and off-line predicted aperture velocity from neural data (solid line, ±2 standard errors) during automatic gripper control, showing that the monkey's cortical population is modulated for observed gripper movement.

PVA, the extraction algorithm used, is dependent on accurate estimates of the recorded units' tuning properties. At the beginning of each day, the tuning properties were estimated in a calibration procedure that did not require the monkey to move its arm. Because motor cortical units modulate their firing rates when the subject watches automatic task performance[21], the assisted task (the same as in the description of training above) was used for calibration. During the first iteration of four trials (one successful trial per target location), the monkey watched the automated performance of reach, grip and retrieval and then received the food. A trial was cancelled if the monkey did not appear to pay attention. Modulation evident during the first iteration was used to get an initial estimate of each unit's tuning properties (Supplementary Methods). During the next iteration, these initial estimates were used by the extraction algorithm to generate a signal that was mixed with the automated control. Tuning parameters were re-estimated at the end of each iteration while gradually decreasing the automated contribution until both arm movement and the gripper were fully controlled by the monkey's cortical activity. An example of the modulation during and after calibration is shown in Fig. 4a. In addition to endpoint movement, in the current study we also used observation-related activity for gripper control (Fig. 4b) in several phases of the training procedure, culminating in its skilled use (Fig. 2e and f).

With this study, we have expanded the capabilities of prosthetic devices through the use of observation-based training and closed-loop cortical control, allowing the use of this four-dimensional anthropomorphic arm in everyday tasks. These concepts can be incorporated into future designs of prostheses for dexterous movement.

METHODS SUMMARY

The timeline of each trial was divided into functional periods (Fig. 1b). A trial began with a piece of food being placed on the presentation device and the device moved to a location within the monkey's workspace to provide a reaching target (Presentation). The monkey often started moving the arm forward slowly before the presentation was complete. When the target was in place, the monkey started a directed reaching movement while simultaneously opening the gripper (Move A). Upon approach, the animal made small homing adjustments to get the endpoint aligned with the target (Home A), and then closed the gripper while actively stabilizing the endpoint position (Loading). If loading was successful, the monkey made a retrieval movement back towards the mouth while keeping the gripper closed (Move B), then made small adjustments to home in on the mouth (Home B) and stabilized the endpoint while using its mouth to unload the food from the gripper (Unloading). A trial was considered successful if the monkey managed to retrieve and eat the presented food. Each trial was followed by an inter-trial period while a new piece of food was prepared for presentation (Inter-trial). During continuous self-feeding, these task periods had no meaning during the execution of the task, but rather were imposed afterwards for purposes of data analysis. In contrast, during training and calibration, a real-time software module kept track of the task periods based on button-presses by a human operator and based on distance of arm endpoint from the tip of the food target presentation device. During training, this real-time delineation of task periods was used so that automated assistance could be applied differently during each task period depending on what aspect of the task the monkey was having difficulty with. During calibration, the delineation of task periods was used so that firing rates collected during each task period could be regressed against appropriate behavioural correlates. Further details on training and calibration are given in Supplementary Methods. Figures 2f and 3c, d are parallel-projection 3D plots. Figure 2g is in perspective-projection.

Received 14 November 2007; accepted 4 April 2008.
Published online 28 May 2008.

1. Georgopoulos, A. P., Kalaska, J. F., Crutcher, M. D., Caminiti, R. & Massey, J. T. in *Dynamic Aspects of Neocortical Function* (eds Edelman, G. M., Gall, W. E. & Cowan, W. M.) 501–524 (Wiley & Sons, New York, 1984).
2. Georgopoulos, A. P., Kettner, R. E. & Schwartz, A. B. Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *J. Neurosci.* **8,** 2928–2937 (1988).
3. Schwartz, A. B. Direct cortical representation of drawing. *Science* **265,** 540–542 (1994).
4. Schwartz, A. B. & Moran, D. W. Motor cortical activity during drawing movements: Population representation during lemniscate tracing. *J. Neurophysiol.* **82,** 2705–2718 (1999).
5. Moran, D. W. & Schwartz, A. B. Motor cortical activity during drawing movements: Population representation during spiral tracing. *J. Neurophysiol.* **82,** 2693–2704 (1999).
6. Moran, D. W. & Schwartz, A. B. Motor cortical representation of speed and direction during reaching. *J. Neurophysiol.* **82,** 2676–2692 (1999).
7. Wessberg, J. *et al.* Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* **408,** 361–365 (2000).
8. Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* **113,** 767–791 (2002).
9. Schwartz, A. B. Cortical neural prosthetics. *Annu. Rev. Neurosci.* **27,** 487–507 (2004).
10. Leuthardt, E. C., Schalk, G., Moran, D. & Ojemann, J. G. The emerging world of motor neuroprosthetics: A neurosurgical perspective. *Neurosurgery* **59,** 1–14 (2006).
11. Schwartz, A. B., Cui, X. T., Weber, D. J. & Moran, D. W. Brain-controlled interfaces: Movement restoration with neural prosthetics. *Neuron* **52,** 205–220 (2006).
12. Kennedy, P. R. & Bakay, R. A. E. Restoration of neural output from a paralyzed patient by a direct brain connection. *Neuroreport* **9,** 1707–1711 (1998).
13. Kennedy, P. R., Bakay, R. A., Moore, M. M., Adams, K. & Goldwaithe, J. Direct control of a computer from the human central nervous system. *IEEE Trans. Rehabil. Eng.* **8,** 198–202 (2000).
14. Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R. & Donoghue, J. P. Instant neural control of a movement signal. *Nature* **416,** 141–142 (2002).
15. Taylor, D. M., Helms Tillery, S. I. & Schwartz, A. B. Direct cortical control of 3D neuroprosthetic devices. *Science* **296,** 1829–1832 (2002).
16. Carmena, J. M. *et al.* Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biol.* **1,** 193–208 (2003).
17. Musallam, S., Corneil, B. D., Greger, B., Scherberger, H. & Andersen, R. A. Cognitive control signals for neural prosthetics. *Science* **305,** 258–262 (2004).
18. Wolpaw, J. R. & McFarland, D. J. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proc. Natl Acad. Sci. USA* **101,** 17849–17854 (2004).
19. Lebedev, M. A. *et al.* Cortical ensemble adaptation to represent velocity of an artificial actuator controlled by a brain-machine interface. *J. Neurosci.* **25,** 4681–4693 (2005).
20. Hochberg, L. R. *et al.* Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* **442,** 164–171 (2006).
21. Wahnoun, R., He, J. & Helms Tillery, S. I. Selection and parameterization of cortical neurons for neuroprosthetic control. *J. Neural Eng.* **3,** 162–171 (2006).
22. Santhanam, G., Ryu, S. I., Yu, B. M., Afshar, A. & Shenoy, K. V. A high-performance brain-computer interface. *Nature* **442,** 195–198 (2006).
23. Chapin, J. K., Moxon, K. A., Markowitz, R. S. & Nicolelis, M. A. L. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neurosci.* **2,** 664–670 (1999).
24. Helms Tillery, S. I., Taylor, D. M. & Schwartz, A. B. The general utility of a neuroprosthetic device under direct cortical control. *Proc. 25th Annu. Int. Conf. IEEE EMBS* **3,** 2043–2046 (2003).
25. Brockwell, A. E., Rojas, A. L. & Kass, R. E. Recursive bayesian decoding of motor cortical signals by particle filtering. *J. Neurophysiol.* **91,** 1899–1907 (2004).
26. Sanchez, J. C., Erdogmus, D., Nicolelis, M. A. L., Wessberg, J. & Principe, J. C. Interpreting spatial and temporal neural activity through a recurrent neural network brain-machine interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* **13,** 213–219 (2005).
27. Yu, B. M. *et al.* Mixture of trajectory models for neural decoding of goal-directed movements. *J. Neurophysiol.* **97,** 3763–3780 (2007).
28. Schwartz, A. B., Taylor, D. M. & Helms Tillery, S. I. Extraction algorithms for cortical control of arm prosthetics. *Curr. Opin. Neurobiol.* **11,** 701–707 (2001).
29. Morasso, P. Spatial control of arm movements. *Exp. Brain Res.* **42,** 223–227 (1981).
30. Soechting, J. F. Effect of target size on spatial and temporal characteristics of a pointing movement in man. *Exp. Brain Res.* **54,** 121–132 (1984).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Clanton and M. Wu for help with software and hardware development, S. Chase for discussions and I. Albrecht for the illustration in Fig. 1a. Support contributed by NIH-NINDS-N01-2-2346 and NIH NSØ5Ø256.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.B.S. (abs21@pitt.edu).

Vol 453|19 June 2008|doi:10.1038/nature06910

LETTERS

Neural substrates of vocalization feedback monitoring in primate auditory cortex

Steven J. Eliades[1] & Xiaoqin Wang[1]

Vocal communication involves both speaking and hearing, often taking place concurrently. Vocal production, including human speech and animal vocalization, poses a number of unique challenges for the auditory system. It is important for the auditory system to monitor external sounds continuously from the acoustic environment during speaking despite the potential for sensory masking by self-generated sounds[1]. It is also essential for the auditory system to monitor feedback of one's own voice. This self-monitoring may play a part in distinguishing between self-generated or externally generated[2,3] auditory inputs and in detecting errors in our vocal production[4]. Previous work in humans[5–10] and other animals[11–13] has demonstrated that the auditory cortex is largely suppressed during speaking or vocalizing. Despite the importance of self-monitoring, the underlying neural mechanisms in the mammalian brain, in particular the role of vocalization-induced suppression, remain virtually unknown. Here we show that neurons in the auditory cortex of marmoset monkeys (*Callithrix jacchus*) are sensitive to auditory feedback during vocal production, and that changes in the feedback alter the coding properties of these neurons. Furthermore, we found that the previously described cortical suppression during vocalization actually increased the sensitivity of these neurons to vocal feedback. This heightened sensitivity to vocal feedback suggests that these neurons may have an important role in auditory self-monitoring.

Vocal communication has an important role in the everyday lives of humans and many other animal species. When we speak, the sound of our voice is both delivered to an intended listener and conducted back to our own ear. Such feedback is a major input to our auditory system during vocal production[1], and is subjected to continuous self-monitoring[4], which requires sensitive detection of vocal feedback changes by neurons in the auditory system. The neural mechanisms underlying vocal feedback monitory are poorly understood.

A small number of previous studies have attempted to investigate the function of the auditory cortex during vocal production. Imaging and neurophysiological studies in humans have shown reduced activity in the auditory cortex during speech production relative to passive listening[5–10]. Similarly, investigations in non-human primates have demonstrated that most auditory cortex neurons exhibit vocalization-induced suppression of neural firing (spontaneous or sound-evoked) during vocal production[11–13]. Previous studies in primates have also shown a smaller subpopulation of auditory cortex neurons that are excited during self-initiated vocalizations. In addition, attenuation of neural signals is also present in the auditory brainstem during vocalization[14], but differs from the vocalization-induced suppression in cortex in that the latter begins several hundred milliseconds before vocal onset[12]. Suppression of auditory cortex neural activity during vocal production contrasts sharply with the typical excitatory responses of cortical neurons in response to the playback of recorded vocalizations that fall into a neuron's receptive field. This suppression is thought to originate from brain regions that initiate and control vocal production. How vocal feedback is encoded during vocal production and the contribution of vocalization-induced suppression to auditory self-monitoring, however, is unclear.

In this study, we examined whether neurons in the auditory cortex were sensitive to auditory feedback during vocal production. Using chronically implanted multi-electrode arrays (Supplementary Fig. 1a, b), we recorded 240 single neurons from the auditory cortices of marmoset monkeys (*Callithrix jacchus*), a highly vocal primate species, while the animals made voluntary, self-initiated vocalizations. By altering the animal's perceived vocal feedback with custom headphones and real-time frequency shifts of ± 2 semitones we found that many auditory cortex neurons were highly sensitive to feedback during vocalization.

Figure 1 illustrates two representative examples of neural responses during vocalization under normal (baseline) or altered feedback conditions. The first neuron (Fig. 1a–c) was suppressed by the animal's own vocalizations under baseline feedback conditions, with a mean response modulation index (RMI) of -0.39. The RMI measures the relative change in firing rate during vocalization as compared with the firing rate before vocalization. This same neuron became strongly excited when the animal vocalized in the presence of $+2$ semitone frequency-shifted feedback (RMI $= 0.70$), as can be seen for multiple vocalizations in the raster (Fig. 1b) and peri-stimulus time histogram (PSTH) (Fig. 1c). As a control, we also tested amplified ($+10$ dB), but not frequency-shifted, feedback and found that it did not change the response from the baseline condition (that is, unaltered feedback). A second neuron (Fig. 1d–f) was excited during normal vocalizations (RMI $= 0.22$) and showed an increase in firing rate under frequency-shifted feedback conditions (RMI $= 0.55$), but not under amplified feedback conditions (RMI $= 0.20$). The firing rate increased significantly in both neurons during frequency-shifted feedback conditions when compared to both unaltered (baseline) and amplified feedback conditions ($P < 0.001$, Kruskal–Wallis analysis of variance (ANOVA)). These examples demonstrate that neurons in the auditory cortex, despite being highly modulated by vocal production, are sensitive to auditory feedback of self-produced vocalizations. This is surprising, particularly for suppressed neurons (for example, Fig. 1a–c) where the vocalization-induced inhibition might be expected to reduce feedback sensitivity.

Overall, neurons suppressed during vocalization, which account for approximately three-quarters of the neurons studied in the auditory cortex[12], exhibited increased activity during frequency-shifted feedback compared with the baseline condition. The average activity of this population of neurons was strongly inhibited during normal vocal production (Fig. 2a). During altered feedback, the firing rate of these neurons increased, but remained suppressed as compared with

[1]Laboratory of Auditory Neurophysiology, Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.

the pre-vocal activity. The second, smaller, population of neurons excited during normal vocalization seemed to be less sensitive to altered feedback (Fig. 2b). The average firing rate of these neurons was slightly reduced in the presence of altered feedback, although the change was much smaller than that observed for suppressed neurons (Fig. 2a). The effect of altered feedback can still be seen when activities of all neurons (suppressed or excited) are averaged together and when different call types are separately analysed (Supplementary Fig. 2).

We analysed the effect of altered feedback on individual neurons within suppressed and excited populations (Fig. 2c and Supplementary Fig. 3). Within these populations, both increases and decreases occurred in neural firing during altered feedback compared to baseline. When plotted against the baseline vocal modulation (unaltered feedback), the changes in RMI due to altered feedback in each neuron confirm the trends shown by population averages (Fig. 2a, b). Altered feedback effects were prominent in neurons with negative baseline RMI values, but not in neurons with positive or near-zero baseline RMIs ($P < 0.001$, Kruskal–Wallis ANOVA). These data indicate that, as a population, suppressed neurons were more sensitive to auditory feedback during vocalization than excited neurons, suggesting that they may have a greater role in vocal self-monitoring.

The presence of feedback-related changes in auditory cortex activity during vocal production raises an important question as to relative contributions of feedback and internal modulations to the observed neural responses. The persistence of reduced firing in suppressed neurons during altered feedback suggests the continued presence of inhibition. A direct comparison of neural responses during vocalization under baseline and frequency-shifted feedback conditions revealed a correlation (Fig. 3a), indicating that feedback combines with, rather than replaces, the underlying vocalization-induced modulation. Across the sampled neurons, both increased and decreased RMIs were observed during altered feedback as compared to the baseline (unaltered) condition (Fig. 3b), but there was an overall bias towards increased neural activity. The directions of frequency shift (+2 versus −2 semitones; Supplementary Fig. 4a) did not change the population responses, and responses were also not different between the two animals (Supplementary Fig. 5).

An alternative explanation for these results is that the differences could have been due to altered vocal production rather than auditory feedback. Auditory cortex neurons are sensitive to natural fluctuations in vocal production[13]. We analysed further the difference in RMI between altered and baseline feedback in a subset of the data

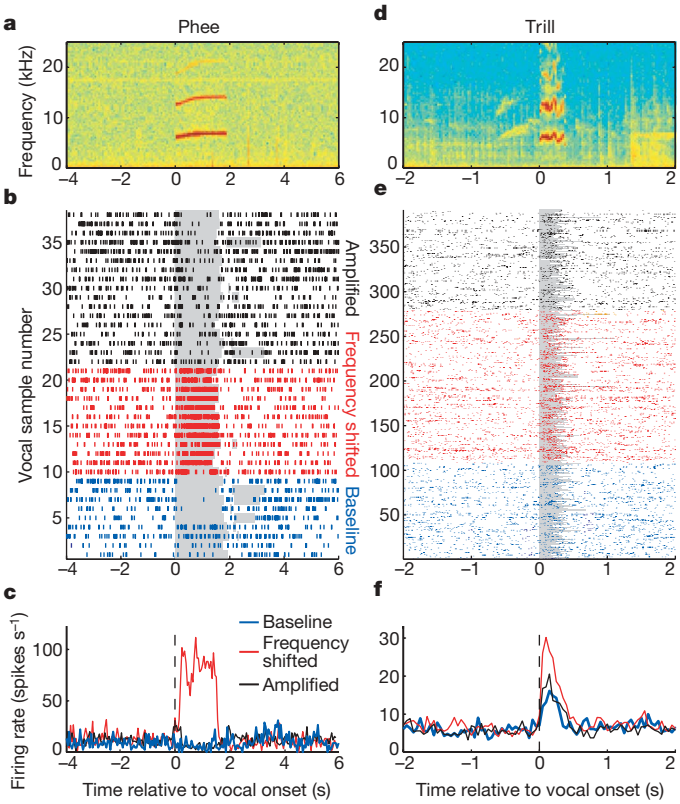


Figure 1 | Examples of vocal suppression and excitation during altered feedback. **a**, Spectrogram of a marmoset phee vocalization. **b**, Raster plot of action potentials before, during and after phees recorded from an auditory cortex neuron that was suppressed during normal vocal production. Shaded areas indicate duration of phees. Neural responses are shown during normal, baseline vocalizations (blue), +2 semitone frequency-shifted feedback (red), and amplified but unshifted feedback (black). Multiple vocalizations and corresponding cortical responses were recorded in each condition. **c**, Peristimulus time histogram (PSTH) illustrating the large increase in firing rate compared to baseline (blue) during frequency-shifted (red), but not amplified (black), feedback. **d**, Spectrogram of a sample trill vocalization. **e**, **f**, Raster plot (**e**) and PSTH (**f**) of an excited neuron whose firing also increased during a +2 semi-tone frequency shift, but not during feedback amplification.

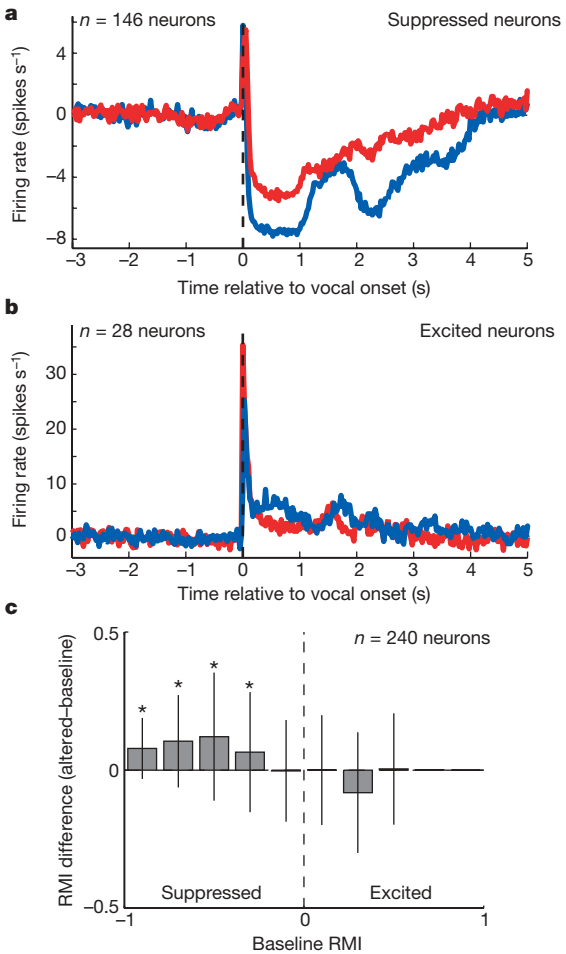


Figure 2 | Feedback effects in suppressed and excited neural populations. **a**, **b**, PSTHs showing average population responses to phee vocalizations in baseline (blue) and frequency-shifted altered feedback (red) conditions. Firing rates during altered feedback were increased in suppressed neurons (**a**; RMI < −0.2), but slightly decreased in excited neurons (**b**; RMI > 0.2). The transient increase in baseline activity (**a**) corresponds to the end of the first phrase of multi-phrased phees. **c**, Relationship between altered feedback effects and baseline vocalization-induced modulation. Differences in the RMI between altered feedback and baseline conditions were calculated for individual neurons ($n = 240$) and averaged for different ranges of baseline RMI. Data are presented as mean values and error bars indicate the s.d. Significant feedback effects are indicated (asterisk, $P < 0.001$, Wilcoxon signed-rank test). Individual data points are shown in Supplementary Fig. 3.

that contained acoustically matched vocalizations in both conditions (Fig. 3c). For each neuron, the average responses for matched and unmatched (acoustically different) vocalizations were calculated during altered feedback conditions. The RMI difference distributions for these three groups were not significantly different ($P > 0.05$, Kruskal–Wallis ANOVA), indicating that changes in neural activity were due to altered acoustic feedback, rather than altered vocal production. Furthermore, this suggests that a previously observed relationship between vocal acoustics and neural modulation[13] was probably due to the auditory feedback rather than variations in the suppressive internal modulations. The existence of auditory feedback-dependent neural responses in auditory cortex is an important observation because, up to this point, it has not been possible to separate out the roles of modulation and feedback. It implies a more complex mechanism, combining both internal modulation and feedback responses, rather than one purely reflecting internal signals.

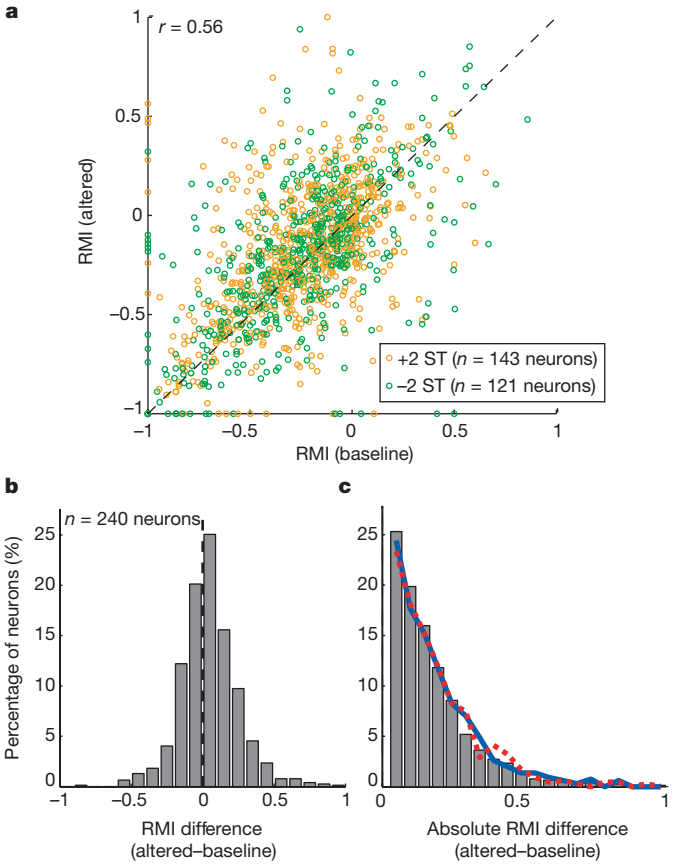


Figure 3 | Population responses to altered feedback. **a**, Scatter plot comparing frequency-shifted RMI with baseline RMI for all neurons and vocalization types. Responses were weakly, but significantly correlated (Spearman rank correlation $r = 0.56$, $P < 0.001$). Positive (orange circles) and negative frequency shifts (green circles) are shown. All three vocalization types studied are included. Points are shown for each vocalization type in each neuron (phee, $n = 197$; trilphee, $n = 162$; trill, $n = 107$). A further breakdown by frequency shift direction (+2 versus −2) and vocalization type is shown as a Supplementary Figure (Supplementary Fig. 4). ST, semitone. **b**, The distribution of the RMI difference between altered and baseline conditions illustrates the presence of both increased and decreased neural activities due to frequency-shifted feedback. The distribution was significantly shifted towards positive RMI differences (mean ± s.d. = 0.05 ± 0.21; $P < 0.001$, Wilcoxon signed-rank test). **c**, The distribution of absolute RMI difference values for all vocalizations (filled bars) is compared with those for which vocalization acoustics during altered feedback either matched (blue line) or did not match (dashed red line) the acoustics of baseline vocalizations. The feedback activity in these three conditions was not significantly affected by acoustic matching ($P > 0.05$, Kruskal–Wallis ANOVA).

A question remains as to the origin of the observed sensitivity to feedback during vocalization. A possible explanation is that that the effects of feedback alteration on vocal responses are due to the shifting of vocal acoustic energy into or out of the auditory receptive fields of auditory cortex neurons. This would suggest that the vocal effects of altered feedback could be predicted from the auditory effects of a similar alteration. We therefore examined the relationship between passive auditory (sensory) responses of auditory cortex neurons and the effects of altered feedback during vocalization. Neurons studied during vocal production were examined further with the playback of vocalizations previously recorded from the same animal. These vocalization stimuli were presented both with and without the same frequency shifts used during vocal feedback manipulations (Supplementary Fig. 6a, b). The difference in neural response was compared to that during vocalization. As seen from the population analysis in Fig. 4a, the auditory and vocal frequency shift effects were uncorrelated, indicating that the neural modulations observed

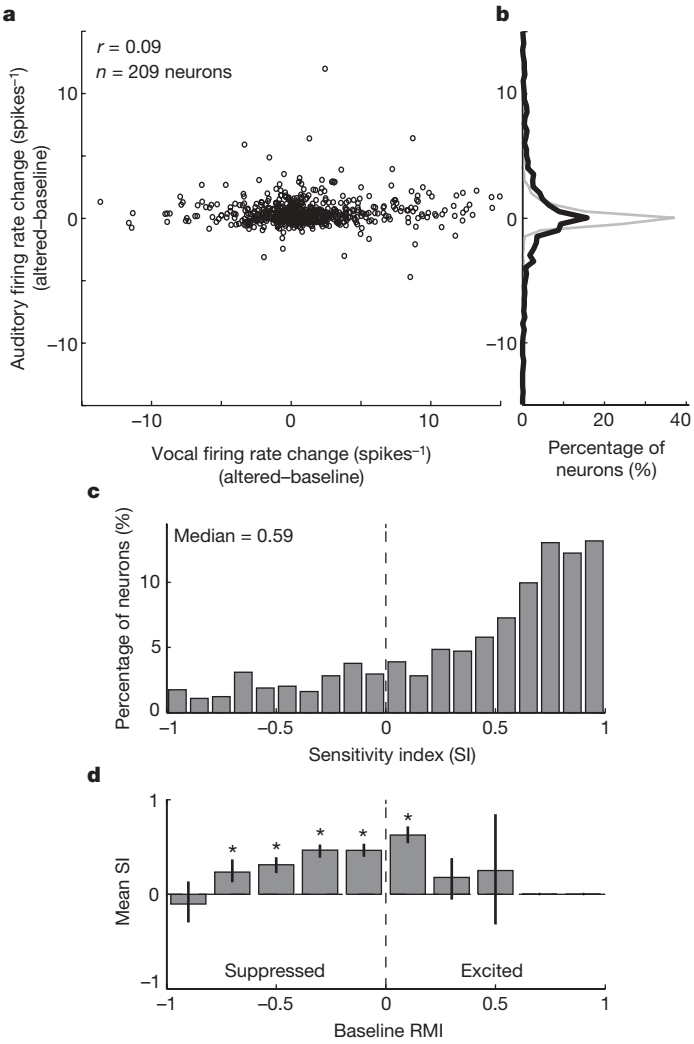


Figure 4 | Auditory responses and feedback sensitivity. **a**, Population scatter plot comparing frequency-shift effects on neural responses during vocal production and auditory playback. The vocal and auditory responses to feedback alterations were not correlated ($r = 0.09$; $P > 0.05$). Distributions of the vocal (black) and auditory (grey) data sets are shown in **b**. **c**, Distribution of the sensitivity index (SI), comparing feedback effects between vocal and auditory for the neurons in **a**, is shown (see Methods). **d**, The relationship between SI and baseline vocalization-induced modulation (measured by RMI) is shown. Most units showed an increase in sensitivity to feedback alteration during vocalization, particularly for suppressed neurons. Error bars represent bootstrapped 95% confidence intervals (asterisk, $P < 0.001$ Wilcoxon signed-rank test).

during altered vocal feedback were not simply due to the auditory (playback) responses of the auditory cortex neurons. Vocal feedback responses were similarly unrelated to the frequency tuning of the neurons (Supplementary Figs 7 and 8). The frequency shift effects were larger (greater firing rate changes) during vocalization than during playback (Fig. 4b). These findings confirm that auditory tuning cannot account for the responses observed during frequency-shifted feedback.

The absence of a clear relationship between auditory and feedback responses suggests that the underlying vocalization-induced modulation may be responsible for changing neural sensitivity during vocal production. We calculated a feedback sensitivity index (SI) for each neuron, which relates the frequency shift effects during vocalization to those during auditory (playback) stimuli. An SI of +1 indicates that a neuron is sensitive to the frequency-shift alteration during vocalization, but not during playback, and an SI of −1 indicates the opposite. The distribution of the SI for the auditory cortical population (Fig. 4c) showed a large concentration towards +1 (median 0.59; $P < 0.001$, Wilcoxon signed-rank test). This indicates that most neurons are more sensitive to frequency shifts during vocalization than predicted from their auditory responses.

To understand this increased feedback sensitivity, we compared the SI with the baseline vocal modulation measured in each neuron (Fig. 4d). The analysis revealed that the largest increases in feedback sensitivity were found in the suppressed and weakly excited neurons. This statistically significant trend ($P < 0.001$, Kruskal–Wallis ANOVA) suggests that vocalization-induced suppression acts to increase, rather than decrease, the sensitivity of auditory cortex neurons to auditory feedback during vocalization. Weakly excited neurons may have been similarly affected because their responses combine elements of both excitation and suppression. The most strongly suppressed neurons failed to show this feedback sensitivity, probably because of the lack of ongoing neural activity during vocalization. Additional analysis of the sensitivity to altered feedback based on a d' (discriminability) measure (Supplementary Fig. 9) demonstrated similar properties to those revealed by the SI analysis (Fig. 4). Population-averaged PSTHs for auditory playback (Supplementary Fig.10) showed, unlike their vocalization counterparts (Fig. 2), larger differences for excited than for suppressed neurons, supporting the notion that the increase in sensitivity to altered feedback is specifically related to vocalization-induced suppression. The observed feedback responses show, for the first time, that one function of vocal suppression is to increase auditory feedback sensitivity during vocalization. Previous work in the cricket cercal (auditory) system has also shown inhibition of auditory responses[15]. However, in contrast to the sensitization during vocalization we have demonstrated in primates, inhibition transiently desensitizes the cricket auditory system to increase sensitivity immediately after stridulation[15].

The effects of vocal feedback alteration or distortion during speaking have been studied previously in the human auditory cortex, with results showing the suppression of the auditory cortex during natural vocalization[5–10] and a small increase in activity during feedback alterations[5,16–18]. The similarity between these results and our current findings suggests that common mechanisms may be shared in the sensory components of both human and non-human primate vocal production. We have previously suggested that the reduced activation of the auditory cortex during speaking seen in human imaging studies may result from a combination of the underlying activity of suppressed and excited neurons[12,13]. It is therefore possible that the altered feedback effects observed in humans may also be due to combinations of increased and decreased activity in individual neurons. Our findings also suggest that the human auditory cortex may exhibit sensitization to auditory feedback during speaking.

How vocalization-induced suppression contributes to the apparent increase in feedback sensitivity remains to be elucidated. One possibility is that the suppression acts to modulate auditory sensitivity non-selectively by scaling the gain of neural responses, thereby magnifying the effects of feedback perturbations. Another possibility is that the modulatory signals (termed corollary discharges or efference copies[19]) contain specific predictions of the expected auditory input (a forward model[20]) that are compared to the actual vocal feedback; the resulting auditory cortex activity represents the deviation from expectancy (error signal). This idea is consistent with self-monitoring for error detection, but conflicts with previously observed changes in AC neural responses that correlate with fluctuations in vocal acoustics in the absence of altered feedback[13]. A final possibility is that modulatory signals induce a transient change in auditory receptive fields during vocalization to better predict acoustic feedback. This might explain why feedback sensitivity exists in neurons whose playback auditory receptive fields do not overlap vocal acoustics. Such changes in receptive field have been observed peri-saccadically in the visuo-motor lateral intraparietal area[21].

Because of the intrinsic variability of marmoset vocal behaviour and the necessity of performing the reported vocal experiments in the marmoset colony, there are several factors that could have affected the apparent increase in feedback sensitivity but cannot be completely controlled for in the present study. For example, differences in background noise between auditory playback (conducted in the sound-proof chamber) and vocalization experiments (conducted in the marmoset colony) were present. Fluctuations in the animal's behavioural state between different experimental conditions may have also been present. Finally, there were potential differences between vocal production and playback because playback stimuli lacked the full diversity of the produced vocalizations. These factors, although unlikely to account completely for the reported observations, need to be kept in mind when interpreting our findings.

Self-monitoring of vocal feedback may have several important functions. In non-human primates, discrimination between self-generated and external sounds may play a part in behaviours where assigning an auditory input as self is important. These include antiphonal calling, an interactive vocal exchange behaviour seen in marmosets and other monkeys[22], and vocal convergence, the tendency of monkeys to match their vocal acoustics to that of their cage-mate[23]. Sensitive monitoring of auditory feedback to detect vocal production errors is also an essential step in feedback-mediated vocal control. Humans constantly monitor their speech and quickly compensate for perceived changes in feedback[24,25], including the frequency-shifted feedback used in these experiments[26]. Feedback-dependant vocal control in non-human primates is less well understood. Monkeys can change the amplitude of their vocalizations when their feedback is disrupted by masking noise[27]. However, there is no published data showing them to exhibit vocal compensation during frequency-shifted feedback. This lack of a direct behavioural correlate for our feedback alterations limits the conclusions that can be drawn from the observed feedback sensitivity in auditory cortex. Nonetheless, a possible role in feedback-mediated vocal control remains an intriguing possibility for future studies, especially as defects in feedback monitoring have been suggested to underlie human communication disorders such as stuttering[28].

METHODS SUMMARY

Two marmoset monkeys were implanted bilaterally with two multi-electrode arrays, one in each auditory cortex. The 16-channel microelectrode arrays (Warp-16, Neuralynx) were adapted from larger multi-electrode designs[29]. Only well-isolated single units were analysed. Neural recordings included both primary auditory cortex and lateral fields and all cortical layers.

Vocalizations were recorded synchronously with neural signals using directional microphones. Experiments were mainly performed with the animals in the setting of the marmoset colony, allowing vocal exchanges between the subject and other animals in the colony. Additional experiments were performed in the laboratory with the animals vocalizing in response to the playback of vocalizations from a loudspeaker. Experiments were conducted with the animal either seated in a primate chair or moving around freely within a cage.

Vocalization feedback was modified in real-time using a digital effects processor and frequency shifts of ±2 semitones. Shifted signals were presented to the

animal through customized headphones at a level ~10 dB SPL (sound pressure level) louder than vocalizations produced by the animal. Auditory control experiments were performed by playing an animal's recorded vocalizations through a loudspeaker at similar amplitudes while the animal sat quietly in a sound-proof chamber.

Neural responses to self-produced vocalizations were quantified using the vocal response modulation index (RMI; see Methods). Sensitivity to feedback alteration was calculated using a sensitivity index (SI; see Methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 November 2007; accepted 13 March 2008.
Published online 4 May 2008.

1. von Békésy, G. The structure of the middle ear and the hearing of one's own voice by bone conduction. *J. Accoust. Soc. Am.* **21,** 217–232 (1949).
2. Johns, L. C. *et al.* Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychol. Med.* **31,** 705–715 (2001).
3. Frith, C. D. *The Cognitive Neuropsychology of Schizophrenia* (Earlbaum Associates, Hillsdale, New Jersey, 1992).
4. Levelt, W. J. Monitoring and self-repair in speech. *Cognition* **14,** 41–104 (1983).
5. Houde, J. F., Nagarajan, S. S., Sekihara, K. & Merzenich, M. M. Modulation of the auditory cortex during speech: an MEG study. *J. Cogn. Neurosci.* **14,** 1125–1138 (2002).
6. Paus, T., Perry, D. W., Zatorre, R. J., Worsley, K. J. & Evans, A. C. Modulation of cerebral blood flow in the human auditory cortex during speech: role of motor-to-sensory discharges. *Eur. J. Neurosci.* **8,** 2236–2246 (1996).
7. Curio, G., Neuloh, G., Numminen, J., Jousmaki, V. & Hari, R. Speaking modifies voice-evoked activity in the human auditory cortex. *Hum. Brain Mapp.* **9,** 183–191 (2000).
8. Ford, J. M. *et al.* Neurophysiological evidence of corollary discharge dysfunction in schizophrenia. *Am. J. Psychiatry* **158,** 2069–2071 (2001).
9. Crone, N. E. *et al.* Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology* **57,** 2045–2053 (2001).
10. Creutzfeldt, O., Ojemann, G. & Lettich, E. Neuronal activity in the human lateral temporal lobe. II. Responses to the subjects own voice. *Exp. Brain Res.* **77,** 476–489 (1989).
11. Müller-Preuss, P. & Ploog, D. Inhibition of auditory cortical neurons during phonation. *Brain Res.* **215,** 61–76 (1981).
12. Eliades, S. J. & Wang, X. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. *J. Neurophysiol.* **89,** 2194–2207 (2003).
13. Eliades, S. J. & Wang, X. Dynamics of auditory-vocal interaction in monkey auditory cortex. *Cereb. Cortex* **15,** 1510–1523 (2005).
14. Suga, N. & Shimozawa, T. Site of neural attenuation of responses to self-vocalized sounds in echolocating bats. *Science* **183,** 1211–1213 (1974).
15. Poulet, J. F. & Hedwig, B. A corollary discharge maintains auditory sensitivity during sound production. *Nature* **418,** 872–876 (2002).
16. Heinks-Maldonado, T. H., Mathalon, D. H., Gray, M. & Ford, J. M. Fine-tuning of auditory cortex during speech production. *Psychophysiology* **42,** 180–190 (2005).
17. Hashimoto, Y. & Sakai, K. L. Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: an fMRI study. *Hum. Brain Mapp.* **20,** 22–28 (2003).
18. Fu, C. H. *et al.* An fMRI study of verbal self-monitoring: neural correlates of auditory verbal feedback. *Cereb. Cortex* **16,** 969–977 (2006).
19. Sperry, R. W. Neural basis of the spontaneous optokinetic responses produced by visual inversion. *J. Comp. Physiol. Psychol.* **43,** 482–489 (1950).
20. Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An internal model for sensorimotor integration. *Science* **269,** 1880–1882 (1995).
21. Duhamel, J. R., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255,** 90–92 (1992).
22. Miller, C. T. & Wang, X. Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. *J. Comp. Physiol. A* **192,** 27–38 (2006).
23. Snowdon, C. T. & Elowson, A. M. Pygmy marmosets modify call structure when paired. *Ethology* **105,** 893–908 (1999).
24. Lane, H. & Tranel, B. The Lombard sign and the role of hearing in speech. *J. Speech Hear. Res.* **14,** 677–709 (1971).
25. Houde, J. F. & Jordan, M. I. Sensorimotor adaptation in speech production. *Science* **279,** 1213–1216 (1998).
26. Burnett, T. A., Freedland, M. B., Larson, C. R. & Hain, T. C. Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am.* **103,** 3153–3161 (1998).
27. Brumm, H., Voss, K., Kollmer, I. & Todt, D. Acoustic communication in noise: regulation of call characteristics in a New World monkey. *J. Exp. Biol.* **207,** 443–448 (2004).
28. Timmons, B. A. & Boudreau, J. P. Auditory feedback as a major factor in stuttering. *J. Speech Hear. Disord.* **37,** 476–484 (1972).
29. Hoffman, K. L. & McNaughton, B. L. Coordinated reactivation of distributed memory traces in primate neocortex. *Science* **297,** 2070–2073 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. McNaughton for sharing implanted multi-electrode recording methods. We acknowledge A. Pistorio for assistance in animal care, M. Melamed for assistance in data collection and C. Miller for his comments on this manuscript. This work was supported by NIH grants to X.W.

Author Contributions S.J.E. and X.W. designed the experiments and co-wrote the paper. S.J.E. carried out the experimental recordings and data analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.J.E. (seliades@jhu.edu) or X.W. (xiaoqin.wang@jhu.edu).

METHODS

Animal preparation and neural recording. Two marmoset monkeys were implanted bilaterally with two multi-electrode arrays (Supplementary Fig. 1a, b), one in each auditory cortex. The 16-channel microelectrode arrays (Warp-16, Neuralynx Inc.) were scaled-down versions of the larger multi-electrode arrays developed for use in studies of rodents and macaque monkeys[29]. Before array placement, animals were implanted with a headcap using procedures previously developed for marmosets[30]. The microelectrodes used were either tungsten or platinum-iridium (impedances 2–4 MΩ). Electrodes were individually moveable using a removable pushing device (Neuralynx).

Neural signals were recorded via a headstage on the animal end of a wire tether, amplified and band-pass filtered, and then digitized onto a personal computer. Neural signals were monitored online to optimize signal quality by electrode movements, and to guide auditory stimulus selection. Action potentials (spikes) were sorted offline using custom software and a principle component-based clustering method. Spikes were classified as either from a single- or multi-unit based on a minimum SNR (signal-to-noise ratio) $>$ 13 dB ($>$4.5:1) and presence of a refractory period. A total of 501 units were recorded during these experiments, of which 240 were later classified as single units. Only single units were included in the data reported here. Neurons were sampled in both hemispheres of the two animals, including primary auditory cortex and lateral fields (lateral belt and parabelt areas) and all cortical layers.

Vocal recording. Acoustic signals were recorded using directional microphones, placed ~20 cm in front of the animals, and digitized synchronously with neural signals. Vocalizations were extracted offline from the recordings and manually classified into established marmoset call types based on spectrograms. Only three of the major vocalization types were included for analysis: phees, trilphees and trills. Vocal experiments were primarily performed with the animals in the setting of the marmoset colony, allowing vocal exchanges between the subject animal and other animals in the colony. Multiple microphones were used to monitor vocalizations produced by the subject and the rest of the colony. Additional experiments were performed in the laboratory with the animal vocally interacting to the playback of vocalizations from a speaker, a behaviour known as antiphonal calling[22]. Animals made a wide variety of vocalizations in the marmoset colony, but only made isolation calls (phee) during antiphonal experiments. Experiments were conducted either with the animal seated in a primate chair but with head restraint removed, or when moving around freely within a small custom-made cage. Wire tethers were used when recording neural signals from free-roaming animals.

Feedback alteration. Vocalization feedback was modified in real-time using a digital effects processor (Yamaha SPX 2000). Frequency shifts of ± 2 semitones were used. This shift magnitude fell within the normal range of marmoset vocal variation. Shifted signals were presented to the animal through customized headphones (Supplementary Fig. 1c, d), modified to attach to the animal's headcap, at a level ~10 dB SPL louder than direct (air-conducted) feedback. Feedback experiments were conducted in a blocked fashion with: (1) an hour of recording baseline (unaltered) vocalizations; (2) an hour of recording with frequency-shifted feedback; and (3) half an hour of recording with amplified, but not frequency-shifted, feedback as a control. More than one frequency shift per session was generally not possible because of time limitations to obtain sufficient vocalizations.

Auditory stimuli. Before vocal recordings, neurons' auditory responses were characterized with the animal seated in a primate chair within a sound-proof chamber. Auditory stimuli were presented free-field through a speaker located 1 m in front of the animal. Centre frequencies of neurons were determined by pure tone or band-pass noise stimuli. Animals were also presented with multiple samples of its own, previously recorded, vocalizations. Frequency-shifted playback stimuli, created from recorded vocalizations using the vocal effects processor, were added back to the original vocal stimuli with an appropriate relative amplitude (+10 dB) and delay (10 ms) to match acoustically the conditions heard during vocal production with altered feedback. Both normal and frequency-shifted stimuli were presented at multiple sound levels, but only those overlapping the produced vocalizations were used for analysis.

Data analysis. Neural responses to self-produced vocalizations were quantified using the vocal response modulation index (RMI). $\mathrm{RMI} = (R_{\mathrm{vocal}} - R_{\mathrm{prevocal}})/(R_{\mathrm{vocal}} + R_{\mathrm{prevocal}})$, where R_{vocal} is the firing rate during vocalization and R_{prevocal} is the firing rate before vocalization. An RMI of −1 indicated complete suppression of neural activity and +1 indicated strongly driven vocalization responses, a low pre-vocal firing rate, or both. The effect of altered feedback on neurons was determined by calculating RMIs for individual vocalizations samples under both baseline (unaltered) and altered-feedback conditions and comparing the average RMI from both conditions. The effects of amplified feedback alone were examined in a subset of data, but found to be negligible in most neurons and were not subjected to further analyses. Population comparisons of feedback effects on suppressed ($\mathrm{RMI} \leq -0.2$) and excited ($\mathrm{RMI} \geq 0.2$) neural populations were made by calculating PSTHs aligned by vocalization onset. Additional analyses compared responses to acoustically matched and unmatched vocalizations. "Matched" vocalizations were those produced during frequency-shifted feedback whose SPL and mean fundamental frequency fell within the 25th–75th percentile range of the vocal acoustics measured for the baseline vocalizations.

Auditory playback effects of feedback alterations were measured by comparing responses to normal and frequency-shifted vocal samples and then compared to feedback effects during vocalization. Frequency-shifted stimuli were combined with normal stimuli to match the conditions during vocal production and altered feedback. A measure of neural sensitivity to auditory feedback alteration, the feedback Sensitivity Index (SI), was calculated as $\mathrm{SI} = (|\Delta \mathrm{FR}_{\mathrm{voc}}| - |\Delta \mathrm{FR}_{\mathrm{aud}}|)/(|\Delta \mathrm{FR}_{\mathrm{voc}}| + |\Delta \mathrm{FR}_{\mathrm{aud}}|)$, where $\Delta \mathrm{FR}_{\mathrm{voc}}$ was the change in firing rate during vocalization between normal and altered feedback, and $\Delta \mathrm{FR}_{\mathrm{aud}}$ was the change in firing rate between normal and frequency-shifted vocal sounds during auditory stimulus playback.

Statistical tests were performed using non-parametric methods, including Wilcoxon rank-sum and signed-rank tests to test differences between distribution medians. Multiple comparisons were performed using Kruskal–Wallis ANOVAs with Bonferroni corrections. Correlation coefficients were carried out using Spearman rank correlations. Confidence intervals were calculated using 200 repetition bootstrapping.

30. Lu, T., Liang, L. & Wang, X. Neural representations of temporally asymmetric stimuli in the auditory cortex of awake primates. *J. Neurophysiol.* **85,** 2364–2380 (2001).

LETTERS

RNA toxicity is a component of ataxin-3 degeneration in *Drosophila*

Ling-Bo Li[1]†, Zhenming Yu[1,2], Xiuyin Teng[1,2] & Nancy M. Bonini[1,2]

Polyglutamine (polyQ) diseases are a class of dominantly inherited neurodegenerative disorders caused by the expansion of a CAG repeat encoding glutamine within the coding region of the respective genes[1]. The molecular and cellular pathways underlying polyQ-induced neurodegeneration are the focus of much research, and it is widely considered that toxic activities of the protein, resulting from the abnormally long polyQ tract, cause pathogenesis[2,3]. Here we provide evidence for a pathogenic role of the CAG repeat RNA in polyQ toxicity using *Drosophila*. In a *Drosophila* screen for modifiers of polyQ degeneration induced by the spinocerebellar ataxia type 3 (SCA3) protein ataxin-3, we isolated an upregulation allele of *muscleblind* (*mbl*), a gene implicated in the RNA toxicity of CUG expansion diseases[4–6]. Further analysis indicated that there may be a toxic role of the RNA in polyQ-induced degeneration. We tested the role of the RNA by altering the CAG repeat sequence to an interrupted CAACAG repeat within the polyQ-encoding region; this dramatically mitigated toxicity. In addition, expression of an untranslated CAG repeat of pathogenic length conferred neuronal degeneration. These studies reveal a role for the RNA in polyQ toxicity, highlighting common components in RNA-based and polyQ-protein-based trinucleotide repeat expansion diseases.

To identify modifiers that add new insight into ataxin-3 pathogenesis, we performed an overexpression enhancer- and promoter-containing *P*-element (*EP*-element) screen with a *Drosophila* model of ataxin-3 (ref. 7) for modifiers of eye degeneration. Seven new *EP*-element insertional mutations were isolated; one of these (B2–E1) dramatically enhanced toxicity, causing severe pigmentation loss and striking collapse of the retina, but had no effect on its own when directed to the eye (Fig. 1a–d). Molecular analysis showed that the B2–E1 insertion was in the promoter of the *mbl* gene, and upregulated gene expression (Supplementary Fig. 2). As Mbl has been implicated as a modifier of CUG repeat RNA toxicity[4–6], these results suggested an unexpected role for *mbl* as a modulator of polyQ protein toxicity.

We confirmed that *mbl* upregulation enhanced polyQ toxicity by generating transgenic flies bearing an *mbl* complementary DNA (cDNA) (MblA, which is implicated in eye and photoreceptor neuron development[8]). As with the original *EP* insertion, flies expressing MblA showed strongly enhanced SCA3trQ61 toxicity, as well as enhanced photoreceptor degeneration when expressed with *rh1-GAL4*, but MblA had no effect on its own (Fig. 1e–h and Supplementary Fig. 3h). B2–E1 and MblA enhanced toxicity of full-length pathogenic forms of ataxin-3 as well as truncated versions, and enhanced the toxicity of pathogenic forms of the Huntington's disease protein (Supplementary Figs 3 and 4). Expression of Mbl, however, did not enhance deleterious eye phenotypes due to compromised chaperone activity or tau (Supplementary Fig. 3d–g). Upregulation of MblA also enhanced the shortened lifespan of flies expressing SCA3trQ78 globally in the nervous system with *elav-GAL4*. Heterozygosity for an *mbl* null allele (mbl^{E27} or mbl^{E2})[8,9] suppressed lifespan reduction (Fig. 1i and Supplementary Table 1), indicating that *mbl* modulated SCA3 toxicity upon both upregulation and with reduced activity.

To reveal insight into potential mechanisms of enhancement, we examined effects of Mbl on polyQ protein and RNA. PolyQ proteins undergo abnormal aggregation, accumulating into nuclear inclusions[2,3]. With the late-onset driver *rh1-GAL4*, polyQ protein accumulates slowly over adult lifespan, allowing sensitive analysis of protein levels and nuclear inclusion formation. MblA increased the level of the polyQ protein, accelerating inclusion formation (Supplementary Fig. 5j, k). Notably, Mbl increased the level of the polyQ RNA as well as protein (Supplementary Fig. 5i), suggesting that Mbl may act on the RNA to enhance polyQ toxicity.

To test whether the interaction with polyQ toxicity was conserved, we determined whether human muscleblind (MBNL1) could also modulate polyQ toxicity in the fly. Flies co-expressing human $MBNL1_{40}$ with pathogenic SCA3trQ78 had dramatically enhanced toxicity with striking loss of pigmentation and accelerated photoreceptor neurodegeneration (Supplementary Fig. 6d, g, i). Previous studies indicated that functional depletion of muscleblind contributes to CUG repeat toxicity, with upregulation of $MBNL1_{40}$ suppressing CUG RNA toxicity[10]. To determine further whether muscleblind modulation of polyQ toxicity shared features with CUG RNA modulation, we asked whether an isoform of MBNL1 that fails to bind CUG repeats in binding assays *in vitro*[6,11] was altered in ability to modulate polyQ toxicity. Transgenics were generated that expressed $MBNL1_{35}$, a form of MBNL1 that lacks exon 4 encoding a domain required for optimal CUG repeat binding[6,11] and conserved in *Drosophila* Mbl (Supplementary Fig. 6a). Flies expressing $MBNL1_{35}$, at levels similar to that of $MBNL1_{40}$ in the presence of SCA3trQ78, had mitigated enhancement, with only mild pigmentation loss and modest loss of photoreceptor neurons (Supplementary Fig. 6). These results indicated that an MBNL1 isoform that is compromised in ability to bind CUG repeat RNA is also compromised in ability to enhance CAG-encoded polyQ toxicity. This suggested that muscleblind may modulate polyQ toxicity at least in part by acting at the level of the RNA, although mechanisms with CUG RNA modulation are likely distinct given the opposite effects in these two disease situations.

The data discussed so far suggested that enhancement by Mbl was associated with an increased level of the RNA with consequent increased level of polyQ protein. Consistent with an effect on the RNA, MBNL1 has been shown to interact with CAG repeat RNA *in vitro* and co-localize with CAG repeat RNA foci in cell culture, suggesting that it can interact with CAG repeat RNA[11,12]. We therefore considered whether the enhanced toxicity observed when

[1]Department of Biology, [2]Howard Hughes Medical Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA. †Present address: Department of Biochemistry, University of Utah, Salt Lake City, Utah 84112-5650, USA.

polyQ RNA levels were increased by Mbl was due simply to an increase in the level of the protein, or whether the CAG repeat RNA itself possessed an inherent toxicity contributing to polyQ disease. The ataxin-3 transgenes used thus far were composed of pure CAG repeat sequences within the glutamine-encoding domain; we therefore tested whether the CAG repeat sequence was important for Mbl enhancement. SCA3 transgenes were generated with a CAG repeat sequence interrupted by CAA codons ($SCA3trQ78_{CAA/G}$). These transgenes are predicted to express a protein of identical amino-acid sequence to the CAG repeat encoding transgenic lines, differing only in the sequence of the RNA within the polyQ-encoding

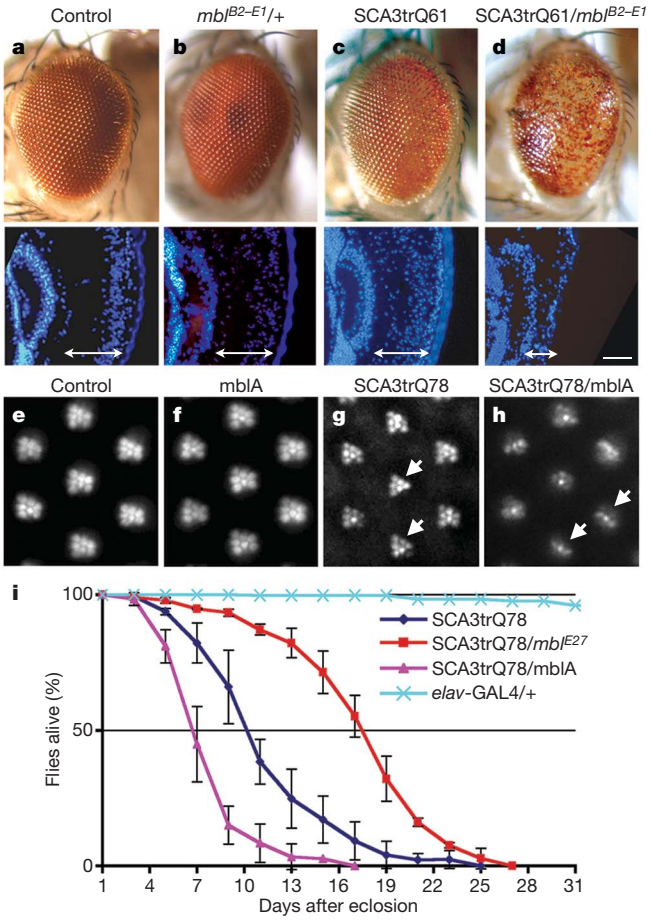


Figure 1 | Upregulation of *mbl* enhances ataxin-3 toxicity. **a**–**d**, External (top) and internal (bottom) retinal cryosections of eyes of 1-day-old flies. **a**, Flies expressing *gmr-GAL4* alone or **b** with mbl^{B2-E1} have normal eye morphology. **c**, Flies expressing SCA3trQ61 have a mild loss of pigmentation, and slightly disrupted internal retinal morphology. **d**, Flies expressing SCA3trQ61 with mbl^{B2-E1} show severe external degeneration and collapse of the retina. Genotypes *w;gmr-GAL4 UAS-SCA3trQ61* in *trans* to **c** *w* or **d** mbl^{B2-E1}. Scale bar in **d**, 5 μm for retinal sections. **e**–**h**, Retinal pseudopupils of 1-day-old flies. **e**, Flies expressing *elav-GAL4* alone or **f** with MblA have a normal pattern of seven photoreceptors per ommatidium. **g**, Flies expressing SCA3trQ78 show mild loss of retinal integrity (arrows), with 5.8 ± 0.4 s.d. photoreceptors per ommatidium ($n = 200$ ommatidia). Genotype *elav-GAL4/+;UAS-SCA3trQ78/+*. **h**, Co-expression of MblA with SCA3trQ78 enhances photoreceptor loss to 3.0 ± 0.5 s.d. ($n = 200$ ommatidia; significant difference from **g**, $P < 0.0001$ (two-tailed unpaired Student's *t*-test)). Genotype *elav-GAL4/+;UAS-mblA/+;UAS-SCA3trQ78/+*. **i**, Neuronal toxicity by lifespan analysis. Upregulation of MblA further shortened the lifespan of SCA3trQ78 flies, whereas downregulation of *mbl* (flies heterozygous for allele mbl^{E27}) extended lifespan ($P < 0.001$, SCA3trQ78 and SCA3trQ78/mbl^{E27}, SCA3trQ78 and SCA3trQ78/mblA, log-rank analysis). Mean ± s.d.; $n = 170$–220 flies for each genotype. Genotypes for SCA3trQ78 and SCA3trQ78/mblA are the same as **g** and **h**, SCA3trQ78/mbl^{E27} is *elav-GAL4/+;*mbl^{E27}*/+;UAS-SCA3trQ78/+*.

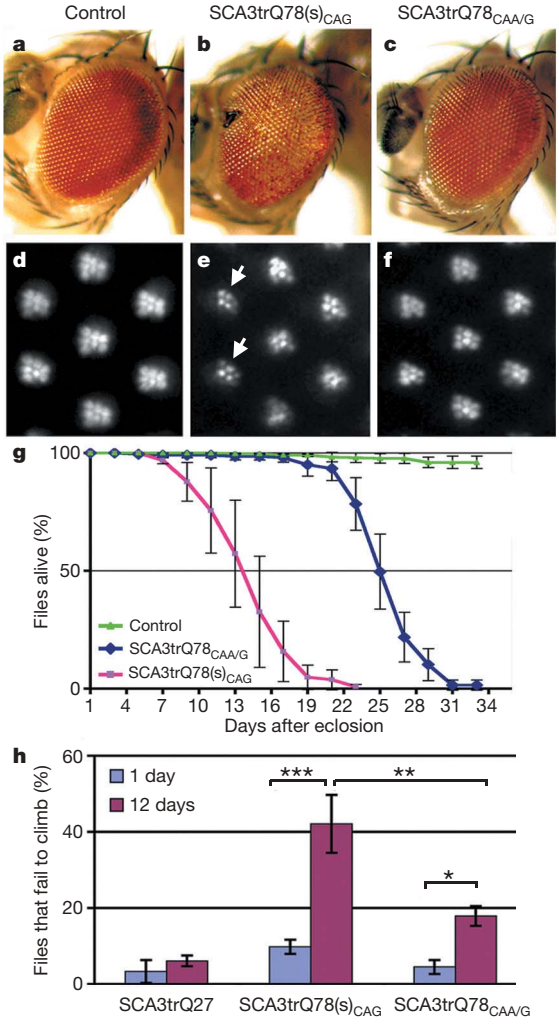


Figure 2 | Interruptions of the CAG repeat mitigate SCA3 protein pathogenesis. **a**–**f**, Flies expressing similar levels of $SCA3trQ78_{CAG}$ and $SCA3trQ78_{CAA/G}$ protein show strikingly different degeneration. **a**–**c**, Flies expressing SCA3 with *gmr-GAL4*. **d**–**f**, Adult-onset photoreceptor retinal degeneration, with expression by *rh1-GAL4* (25-day-old flies). **a**, **d**, Controls expressing non-pathogenic SCA3trQ27 protein have normal eye structure. Genotype *UAS-SCA3trQ27* in *trans* to **a** *gmr-GAL4* or **d** *rh1-GAL4*. **b**, **e**, Flies expressing $SCA3trQ78(s)_{CAG}$ show **b** severe retinal degeneration and **e** (arrows) striking loss of rhabdomeres (4.6 ± 0.6 s.d. photoreceptors per ommatidium). *UAS-SCA3trQ78(s)$_{CAG}$* in *trans* to **b** *gmr-GAL4* and **e** *rh1-GAL4*. **c**, **f**, Flies expressing $SCA3trQ78_{CAA/G}$ at the same level show mild degeneration, with **c** normal external eye morphology and **f** mild photoreceptor loss (6.5 ± 0.2 s.d. photoreceptors per ommatidium, statistically significant from **e**, $P < 0.001$, two-tailed unpaired Student's *t*-test). Genotype *UAS-SCA3trQ78$_{CAA/G}$* in *trans* to **c** *gmr-GAL4* or **f** *rh1-GAL4*. **g**, Neuronal toxicity by lifespan analysis. Flies expressing $SCA3trQ78(s)_{CAG}$ have a strikingly shorter lifespan than flies expressing $SCA3trQ78_{CAA/G}$ at the same level ($P < 0.001$, log-rank test). Mean ± s.d., $n = 150$–200 flies for each. Genotypes: *elav-GAL4* in *trans* to *UAS-SCA3trQ78(s)$_{CAG}$* or *UAS-SCA3trQ78$_{CAA/G}$*. **h**, Climbing behaviour with age. At 1 day, both $SCA3trQ78_{CAG}$ and $SCA3trQ78_{CAA/G}$ flies show normal climbing compared with SCA3trQ27 control flies, with only about 5% failing to climb with agitation (mean ± s.d., $n = 120$–200 flies in total). $SCA3trQ78_{CAG}$ flies show more progressive climbing defects, such that at 12 days, 42.1 ± 7.6% of the flies fail to climb (mean ± s.d., $n = 180$). In contrast, only 17.9 ± 2.6% of $SCA3trQ78_{CAA/G}$ flies fail to climb at 12 days (mean ± s.d, $n = 200$). Genotypes as in ***h***. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ (two-way analysis of variance).

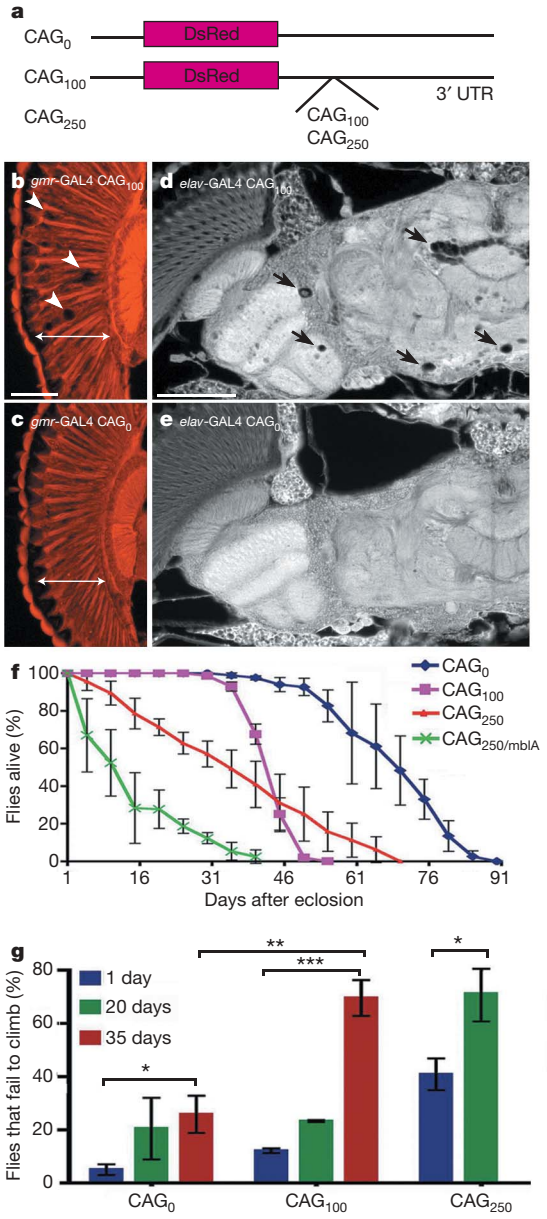


Figure 3 | Untranslated CAG repeats induce progressive neural dysfunction. **a**, Constructs with untranslated CAG repeats within the 3′ untranslated region of a transgene encoding a control protein DsRed. **b**–**e**, Flies expressing untranslated CAG repeats show neuronal degeneration. **b**, **c**, Flies expressing CAG_{100} in the eye with *gmr-GAL4* showed loss of retinal integrity (arrows). Paraffin sections of 1-day-old flies, genotypes: **b**, *gmr-GAL4/UAS-CAG$_{100}$* (4×); **c**, *gmr-GAL4/UAS-CAG$_0$*. **d**, **e**, Flies expressing CAG_{100} showed progressive brain degeneration with vacuoles in the brain (arrows). Paraffin sections of 35-day-old flies, genotypes: **d**, *elav-GAL4/UAS-CAG$_{100}$* (5×) and **e** *elav-GAL4/UAS-CAG$_0$*. Scale bar in **b**, 5 μm for **b**, **c**; scale bar in **d**, 10 μm for **d**, **e**. **f**, **g**, Expression of untranslated CAG repeats induces length-dependent, progressive neural dysfunction. **f**, Neurotoxicity by lifespan analysis. Flies expressing untranslated CAG repeats show length-dependent reduced lifespan. Differences in lifespan of flies expressing CAG_0, CAG_{100} and CAG_{250}, and CAG_{250}/mblA, are significantly different at $P < 0.001$ (log-rank analysis). Mean ± s.d., $n = 250$, 260, 300, 100 flies, respectively. *elav-GAL4* in *trans* to (CAG_0) *UAS-CAG$_0$*, (CAG_{100}) *UAS-CAG$_{100}$* (5×), ($CAG_{250/mblA}$) is *elav-GAL4/+;UAS-mblA/+;UAS-CAG$_{250}$/+*. (CAG_{250}) *UAS-CAG$_{250}$*. **g**, Climbing ability with age. Flies expressing CAG_0 showed normal climbing defects with age ($^*P < 0.05$). Flies expressing CAG_{100} had mild climbing defects at 20 days, which strikingly degenerated by 35 days ($^{***}P < 0.001$ compared with 1 day; $^{**}P < 0.01$ compared with 35 days CAG_0). CAG_{250} flies had moderate climbing defects at 1 day, which were strikingly worse by 20 days ($^*P < 0.05$). Mean ± s.d., 100–200 flies per time point for each genotype in each experiment, two-way analysis of variance. Genotypes as in **f**.

region (Supplementary Fig. 7a). Lines were selected that expressed the protein at levels similar to those of CAG repeat transgenic lines (Supplementary Fig. 7b, c and Supplementary Table 2).

Analysis of the toxicity in flies expressing similar levels of $SCA3trQ78_{CAA/G}$ and the pure CAG-encoded protein ($SCA3trQ78_{CAG}$) showed that the two transgenes conferred strikingly different degrees of degeneration. Whereas $SCA3trQ78(s)_{CAG}$ flies had strong loss of pigmentation and photoreceptor neurons (4.6 ± 0.6 photoreceptors per ommatidium), $SCA3trQ78_{CAA/G}$ flies expressing identical levels of protein had only mild pigmentation loss and minimal degeneration (6.5 ± 0.2 photoreceptors per ommatidium) (Fig. 2a–f and Supplementary Fig. 8). Mitigation of toxicity also occurred with expression globally in the nervous system with *elav-GAL4*: $SCA3trQ78_{CAA/G}$ flies lived longer, with milder progression of neuronal dysfunction (Fig. 2g, h). These findings were confirmed with multiple independent $SCA3trQ78_{CAA/G}$ and $SCA3trQ78_{CAG}$ transgenic lines (Supplementary Table 2). We also analysed flies expressing full-length ataxin-3. These studies showed similar mitigation of toxicity when the CAG repeat sequence was replaced by an interrupted CAA/G sequence within ataxin-3 encoding transgene (Supplementary Fig. 9). These results indicated that toxicity of pathogenic ataxin-3 is mitigated upon interruption of the CAG repeat.

Western blots and immunohistochemistry confirmed that flies expressing pure CAG or CAA/G interrupted transgenes for truncated or full-length ataxin-3 expressed similar levels of polyQ protein. In both, protein was initially diffuse, then accumulated into inclusions of similar size and intensity over time, despite the differential toxicity (Supplementary Figs 7c and 9b, c). The $SCA3trQ78_{CAA/G}$ flies, however, had lower steady-state levels of the transgene-encoding RNA (50 ± 10% of that of $SCA3trQ78(s)_{CAG}$ flies), indicating that interruption of the CAG repeat sequence affected the ratio of SCA3trQ78 transcript to protein (Supplementary Fig. 7b). We also tested whether interrupting the RNA sequence affected the interaction with Mbl. Whereas MblA normally dramatically enhanced polyQ toxicity, it had only a modest effect on flies expressing $SCA3trQ78_{CAA/G}$, with minimal pigmentation and photoreceptor loss (Supplementary Fig. 10). Results were similar with co-expression of human $MBNL1_{40}$ (data not shown). Moreover, whereas upregulation of MblA increased the level of the CAG repeat RNA by 1.6- ± 0.2-fold, it did not affect the level of $SCA3trQ78_{CAA/G}$ RNA (Supplementary Fig. 10i). This result suggested that enhancement of ataxin-3 toxicity by MblA correlated with increased levels of CAG repeat RNA.

These results suggested that a pathogenic length CAG repeat RNA may contribute to polyQ toxicity beyond coding for a pathogenic polyQ protein. This raised the possibility that expression of RNAs containing non-coding CAG repeats of pathogenic length may, on their own, be deleterious. We therefore generated flies bearing a pathogenic length CAG repeat within the 3′ untranslated region of the DsRed open reading frame (Fig. 3a). Repeats of about 100 and 250 CAGs were examined, the latter being within the upper limit of expansions in polyQ diseases like SCA2 and SCA7 (ref. 2). Transgenic lines were selected or combined that expressed the RNA at levels similar to that of $SCA3trQ78_{CAG}$ protein-encoding transgenes (Supplementary Fig. 11a). We confirmed that the non-coding CAG repeat transgenics did not express a CAG repeat containing protein or generate anti-sense CUG transcripts (Supplementary Fig. 11 and data not shown).

Our study revealed that expression of untranslated CAG repeats caused late-onset loss of neuronal integrity when directed to the eye and nervous system. Flies expressing CAG repeats (CAG_{100} or CAG_{250}) in the eye showed internal loss of retinal tissue, with holes within the retina (Fig. 3b, c and data not shown). When expressed globally in neurons with *elav-GAL4*, flies bearing CAG_{100} had normal brain morphology at 1 day (data not shown), but underwent striking degeneration of the brain with time, and died early with progressive loss of climbing compared to controls with CAG_0 (Fig. 3d–g). CAG

repeat-induced neuronal dysfunction was repeat-length dependent, with flies expressing CAG_{250} at lower levels than CAG_{100} showing more severe defects (Fig. 3f, g). MblA also enhanced toxicity of the untranslated CAG RNA (Fig. 3f), and could increase the level of the transcript (Supplementary Fig. 12), indicating that Mbl enhancement of neurotoxicity does not require polyQ protein. We further determined that a long, untranslated, but interrupted RNA construct ($CAA/G_{100\ or\ 262}$), when expressed at levels comparable to CAG_{250} lines, was not toxic (Supplementary Fig. 13). These results confirmed that pathogenic length CAG repeat RNAs can induce neuronal degeneration independent of coding for a toxic polyQ protein.

We next examined whether untranslated CAG repeats induce toxicity through similar mechanisms as other non-coding trinucleotide repeat diseases; namely like CTG repeat sequences. We examined whether CAG repeat RNA formed RNA foci or affected RNA splicing, which are hallmarks of CUG repeat diseases[13,14]. We found that the CAG repeat RNA (CAG_{250}) could form RNA foci, although the foci were present in a limited number of cells and were small compared with those formed by CUG repeat RNA (Supplementary Fig. 14). There was no effect on the splicing pattern of a reporter construct[15] when expressed in the fly (Supplementary Fig. 15). These results suggest that untranslated CAG repeat RNA induces neurodegeneration through mechanisms different from those of CUG repeat RNA.

Taken together, our results suggest that CAG repeat RNAs may contribute a toxic component in polyQ disease, inducing neuronal dysfunction and progressive degeneration (Supplementary Fig. 1). In support of a role of the RNA in polyQ disease are our findings that interruptions of the CAG repeat within protein coding transgenes mitigate toxicity, and that an untranslated CAG repeat RNA can cause toxicity on its own. Although previous studies suggest a minimal role of the RNA in polyQ toxicity[16], raw polyQ domains used in such studies induce strikingly more severe toxicity than polyQ domains within a host protein context[17]; this may mask toxic contributions of the RNA. We find that flies expressing pathogenic SCA3 polyQ proteins encoded by CAA/G repeat transgenes can still develop degenerative phenotypes, but they are less severe and progress more slowly than those of $SCA3trQ78_{CAG}$ flies with similar levels of protein. We therefore suggest that the RNA contributes to a shift of the toxicity curve, rather than being solely responsible for toxicity. Host protein context clearly has a crucial role in polyQ diseases: indeed, a change of a single amino acid outside the polyQ stretch can substantially diminish pathogenesis, despite expression of an expanded CAG repeat RNA[18–20]. Within the protein context of ataxin-3, including truncated forms of the protein that occur in disease[21], effects of the CAG repeat RNA may be particularly apparent. Despite the striking neural dysfunction and degeneration caused by CAG repeat RNA, we did not observe effects on external eye morphology. This is consistent with a previous study that examined only the external eye[16], and is reminiscent of toxicity of the full-length ataxin-3 disease gene in *Drosophila*[22]; these findings suggest a selective sensitivity of neurons to the RNA toxicity.

The pathogenic effects of the CAG repeat RNA may extend to interactions with other cellular factors, as revealed by Mbl. Human MBNL1 was initially identified as a CUG RNA-binding protein[6], although it has been shown to also interact with CAG repeat RNA[11,12]. Our studies reveal that *mbl* modulation of polyQ toxicity is affected by the polyQ-encoding CAG sequence of ataxin-3, as well as by the domain of MBNL1 required for optimal CUG repeat RNA binding. Comparison of flies expressing the same steady-state level of $polyQ_{CAG}$ RNA in the presence or absence of added Mbl indicated that polyQ toxicity with Mbl was more severe (Supplementary Fig. 16), implying that Mbl enhancement is not exclusively due to an increase in the level of the RNA transcript. These results suggest that muscleblind may act at least in part at the level of the CAG repeat RNA in the animal *in vivo*. Additional studies are required to determine the role of muscleblind in human polyQ disease.

Many human neurological diseases are caused by nucleotide repeat expansions[2]. Although the polyQ diseases have typically been viewed as a separate class due to a protein toxicity, other repeat-expansion diseases like DM1, fragile X-associated ataxia and SCA8 are associated with deleterious actions of an expanded repeat RNA[23–25]. Data with DM1 and SCA8—both CUG repeat RNA expansion diseases—also indicate more complex mechanisms that include the production of anti-sense CAG repeat RNAs which may act at the level of the RNA and/or generation of polyQ protein[26,27]. Our findings that the CAG repeat RNA in the ataxin-3 situation may confer pathogenicity highlights additional aspects of these two classes of disease that may be shared.

METHODS SUMMARY

General fly lines were from the Bloomington *Drosophila* Stock Center. SCA3trQ78 fly lines have been described previously[7]. The $SCA3trQ78_{CAA/G}$ and $SCA3nQ81_{CAA/G}$ constructs were generated from $MJDQ82_{CAA/G}$ (ref. 28), using the QuikChange Site-Directed Mutagenesis kit (Stratagene). The MblA cDNA was generated by polymerase chain reaction with reverse transcription (RT–PCR) from wild-type flies. $MBNL1_{35}$ was generated from $MBNL1_{40}$ (ref. 29) using site-directed mutagenesis. CAG_{100} and CAG_{250} repeats were inserted into the 3′ untranslated region of DsRed2 gene (Clontech) in the pUAST vector. Non-coding CAG repeats were sized by genescan analysis[30] with primers: forward 5′-CGTGGAGCAGTACGAGCGCAC-3′; reverse 5′-AGGTTCCTTCACAAAGATCCTC-3′. This analysis showed that the CAG_{100} lines had repeats that were moderately unstable over time, and at the end of the analysis had repeats of about 102–106 (although several lines had flies with repeats of about 85–90, and one line contracted to 17). Multiple lines were combined to achieve levels of the RNA comparable to that of the SCA3trQ78 lines, and are noted (for example, 5× means five independent transgenic lines were combined for expression). The CAG_{250} lines had repeat lengths of about 240–270, and an expression level of about 25% that of the SCA3trQ78 lines. Techniques for western and northern blots were standard; details are described in Methods. Photoreceptor counts were performed using the corneal optical neutralization technique. For each data point, 10 ommatidia per eye and 15–20 eyes per genotype were scored; for at least one experiment of each type, genotypes were scored in a blinded manner. The climbing ability of the flies was analysed by negative geotaxis. Groups of about 20 females were anaesthetized and placed in a vertical plastic column. After 30 min recovery, flies were banged to the bottom, then scored for climbing ability as the percentage of flies remaining at the bottom (less than 2 cm) at 25 s. Three trials were performed at 3 min intervals in each experiment. One hundred to two hundred flies were tested per genotype. Additional details are given in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 November 2007; accepted 12 February 2008.
Published online 30 April 2008.

1. Orr, H. T. & Zoghbi, H. Y. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30,** 575–621 (2007).
2. Gatchel, J. R. & Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Rev. Genet.* **6,** 743–755 (2005).
3. Ross, C. A. & Poirier, M. A. What is the role of protein aggregation in neurodegeneration? *Nature Rev. Mol. Cell Biol.* **6,** 891–898 (2005).
4. Jiang, H., Mankodi, A., Swanson, M. S., Moxley, R. T. & Thornton, C. A. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum. Mol. Genet.* **13,** 3079–3088 (2004).
5. Mankodi, A. *et al.* Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. *Hum. Mol. Genet.* **10,** 2165–2170 (2001).
6. Miller, J. W. *et al.* Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J.* **19,** 4439–4448 (2000).
7. Warrick, J. M. *et al.* Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in *Drosophila. Cell* **93,** 939–949 (1998).
8. Begemann, G. *et al. muscleblind*, a gene required for photoreceptor differentiation in *Drosophila*, encodes novel nuclear Cys_3His-type zinc-finger-containing proteins. *Development* **124,** 4321–4331 (1997).
9. Artero, R. *et al.* The *muscleblind* gene participates in the organization of Z-bands and epidermal attachments of *Drosophila* muscles and is regulated by Dmef2. *Dev. Biol.* **195,** 131–143 (1998).
10. Kanadia, R. N. *et al.* Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy. *Proc. Natl Acad. Sci. USA* **103,** 11748–11753 (2006).

11. Kino, Y. *et al.* Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats. *Hum. Mol. Genet.* **13,** 495–507 (2004).
12. Ho, T. H. *et al.* Colocalization of muscleblind with RNA foci is separable from mis-regulation of alternative splicing in myotonic dystrophy. *J. Cell Sci.* **118,** 2923–2933 (2005).
13. Ranum, L. P. & Cooper, T. A. RNA-mediated neuromuscular disorders. *Annu. Rev. Neurosci.* **29,** 259–277 (2006).
14. Wheeler, T. M. & Thornton, C. A. Myotonic dystrophy: RNA-mediated muscle disease. *Curr. Opin. Neurol.* **20,** 572–576 (2007).
15. Philips, A. V., Timchenko, L. T. & Cooper, T. A. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science* **280,** 737–741 (1998).
16. McLeod, C. J., O'Keefe, L. V. & Richards, R. I. The pathogenic agent in *Drosophila* models of 'polyglutamine' diseases. *Hum. Mol. Genet.* **14,** 1041–1048 (2005).
17. Marsh, J. L. *et al.* Expanded polyglutamine peptides alone are intrinsically cytotoxic and cause neurodegeneration in *Drosophila*. *Hum. Mol. Genet.* **9,** 13–25 (2000).
18. Emamian, E. S. *et al.* Serine 776 of ataxin-1 is critical for polyglutamine-induced disease in SCA1 transgenic mice. *Neuron* **38,** 375–387 (2003).
19. Graham, R. K. *et al.* Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell* **125,** 1179–1191 (2006).
20. Klement, I. A. *et al.* Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in SCA1 transgenic mice. *Cell* **95,** 41–53 (1998).
21. Goti, D. *et al.* A mutant ataxin-3 putative-cleavage fragment in brains of Machado-Joseph disease patients and transgenic mice is cytotoxic above a critical concentration. *J. Neurosci.* **24,** 10266–10279 (2004).
22. Warrick, J. M. *et al.* Ataxin-3 suppresses polyglutamine neurodegeneration in *Drosophila* by a ubiquitin-associated mechanism. *Mol. Cell* **18,** 37–48 (2005).
23. Jin, P. *et al.* RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in *Drosophila*. *Neuron* **39,** 739–747 (2003).
24. Ranum, L. P. & Day, J. W. Pathogenic RNA repeats: an expanding role in genetic disease. *Trends Genet.* **20,** 506–512 (2004).
25. Tapscott, S. J. & Thornton, C. A. Reconstructing myotonic dystrophy. *Science* **293,** 816–817 (2001).
26. Cho, D. H. *et al.* Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol. Cell* **20,** 483–489 (2005).
27. Moseley, M. L. *et al.* Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nature Genet.* **38,** 758–769 (2006).
28. Osherovich, L. Z. & Weissman, J. S. Multiple Gln/Asn-rich prion domains confer susceptibility to induction of the yeast [PSI(+)] prion. *Cell* **106,** 183–194 (2001).
29. Garcia-Casado, M. Z., Artero, R. D., Paricio, N., Terol, J. & Perez-Alonso, M. Generation of GAL4-responsive muscleblind constructs. *Genesis* **34,** 111–114 (2002).
30. Jung, J. & Bonini, N. CREB-binding protein modulates repeat instability in a *Drosophila* model for polyQ disease. *Science* **315,** 1857–1859 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Cashmore, D. Lessing and R. Pittman for comments, J. Weissman, C. Thornton, M. Baylies, R. Artero and the Developmental Studies Hybridoma Bank (supported by the National Institute of Child Health and Human Development and the University of Iowa) for reagents. These studies were supported by the National Institute of Neurological Disorders and Stroke (to N.M.B.). N.M.B. is an Investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.M.B. (nbonini@sas.upenn.edu).

METHODS

Fly lines. Flies were grown at 25 °C on standard medium, unless otherwise indicated. The MblA cDNA was generated by RT–PCR from wild-type flies, with primers: forward 5′-CCGGCCAGATCTACGATGGCCAACGTTGTCAATAT-GAAC-3′; reverse 5′-CCGGCCGTCGACAATTGACTTCATTGGATACAT-AAAC-3′. Photoreceptor counts were performed using the corneal optical neutralization technique, with details as in Methods Summary. The climbing ability of the flies was analysed by negative geotaxis, as described in Methods Summary.

Western and northern blots. Primary antibodies were anti-HA (3F10, 1:500, Roche rat-HRP conjugate), anti-tubulin (E7, 1:10,000, Developmental Studies Hybridoma Bank), anti-MBNL1 (1:4,000) (courtesy of C. Thornton) anti-DsRed (1:400, Clontech, anti-rabbit) and anti-polyQ (1C2, 1:250, Chemicon, anti-mouse). HRP-conjugated secondary anti-mouse (1:4,000, Chemicon) and anti-rabbit antibodies (1:4,000, Zymed) were used with ECL+ reagent (Amersham). For northern blots, total RNA was isolated from adult heads using Trizol (Invitrogen). Probes for northern blots (to SCA3, SV40 and rp49) were labelled using the High Prime kit (Roche). Primers to generate a probe to the SCA3 transgenes were: forward, 5′-CTATCAGGACAGAGTTCACAT-3′; reverse, 5′-CAGATAAAGTGTGAAGGTAGC-3′. Primers for the SV40 common region of UAS-transgenes were: forward 5′-TGTGGTGTGACATAATTGG-ACA-3′; reverse 5′-AGATGGCATTTCTTCTGAGCA-3′.

Real-time PCR. Total RNA from 20 heads of 0- to 4-day-old flies was extracted using RNAeazy kit (Qiagen) and treated with the TURBO DNA-free kit (Ambion). cDNA was synthesized in a 20-μl reaction volume from 0.2 μg of total RNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). cDNA (1 μl) was used as the template in a 20-μl reaction volume diluted from the Power SYBR Green PCR Master Mix Kit or the TaqMan Fast Universal PCR Master Mix (both from Applied Biosystems). Taqman probes and primers were designed using the software Primer Express version 2.0 (Applied Biosystems). Real-time PCR was performed in triplet or quadruplicate using a 7500 Fast Real-Time PCR System (Applied Biosystems). Data were analysed using either the ΔΔCt method (for the splicing assay and DsRed transgene levels) or the standard curve method (for *mbl* levels). Endogenous controls were either rp49 or Gal4. The entire experiments were repeated two to four times on independent RNA preparations.

***In situ* hybridization.** The body wall muscles of third-instar larvae were fixed in 4% formaldehyde/PBS solution for 30 min at 4 °C, pre-hybridized for 10 min, hybridized with probe (2 ng μl^{-1}) for 2 h at 45 °C in buffer (30% formamide, 2× SSC, 0.02% bovine serum albumin, yeast tRNA (1 mg ml^{-1}), 200 mM vanadate), then washed in 30% formamide, 2× SCC for 30 min at 45 °C followed by 1× SCC for 30 min at 22 °C. For nuclear counterstaining, samples were treated with 10 μg μl^{-1} RNaseA (Qiagen) in 2× SSC at 37 °C for 30 min, incubated in 500 nM propodium iodide (Invitrogen) in 2× SSC for 5 min, and then washed three times in 2× SSC. The probes were fluorescein 5′ end-labelled $(CAG)_7$ or $(CUG)_7$ 2-O-methyl RNA 20-nucleotide oligomers (IDT).

LETTERS

Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype

Helene R. McMurray[1]*, Erik R. Sampson[1]*, George Compitello[1]*, Conan Kinsey[1]*, Laurel Newman[1], Bradley Smith[1], Shaw-Ree Chen[1], Lev Klebanov[2,4], Peter Salzman[2], Andrei Yakovlev[2,3]‡ & Hartmut Land[1,3]

Understanding the molecular underpinnings of cancer is of critical importance to the development of targeted intervention strategies. Identification of such targets, however, is notoriously difficult and unpredictable. Malignant cell transformation requires the cooperation of a few oncogenic mutations that cause substantial reorganization of many cell features[1] and induce complex changes in gene expression patterns[2–6]. Genes critical to this multifaceted cellular phenotype have therefore only been identified after signalling pathway analysis[7–10] or on an *ad hoc* basis[4,11–14]. Our observations that cell transformation by cooperating oncogenic lesions depends on synergistic modulation of downstream signalling circuitry[15–17] suggest that malignant transformation is a highly cooperative process, involving synergy at multiple levels of regulation, including gene expression. Here we show that a large proportion of genes controlled synergistically by loss-of-function p53 and Ras activation are critical to the malignant state of murine and human colon cells. Notably, 14 out of 24 'cooperation response genes' were found to contribute to tumour formation in gene perturbation experiments. In contrast, only 1 in 14 perturbations of the genes responding in a non-synergistic manner had a similar effect. Synergistic control of gene expression by oncogenic mutations thus emerges as an underlying key to malignancy, and provides an attractive rationale for identifying intervention targets in gene networks downstream of oncogenic gain- and loss-of-function mutations.

To identify genes regulated synergistically by cooperating oncogenic mutations at genomic scale, we compared messenger RNA expression profiles of young adult murine colon (YAMC) cells with those of YAMC cells expressing mutant $p53^{175H}$ (mp53), activated H-RasV12 (Ras) or both mutant proteins together (mp53/Ras)[17] using Affymetrix microarrays. Using a stepwise procedure, we first identified 538 differentially expressed genes between mp53/Ras and YAMC control cells with a statistical cutoff at $P < 0.01$ (N-test, Westfall–Young adjusted). A further subset of 95 annotated genes that responded synergistically (28 upregulated, 67 downregulated) to the combination of mutant p53 and Ras proteins (termed 'cooperation response genes', CRGs) was then determined using a synergy score as described in Methods (Fig. 1, Supplementary Table 1 and Supplementary File 1). Expression values and synergy scores for the CRGs derived from TaqMan low-density quantitative polymerase chain reaction (qPCR) array data showed strong positive correlation with the values for the same genes obtained from microarray analysis (Supplementary Figs 1 and 2, Supplementary Table 2 and Supplementary File 2). CRG identification was therefore confirmed by independent methods, with final CRG selection based on microarray data due to the higher sample replication in this dataset.

CRGs encode proteins which are involved in the regulation of cell signalling, transcription, apoptosis, metabolism, transport or adhesion (Fig. 2a, b and Supplementary Table 1), and a large proportion are misexpressed in human cancer. For 47 out of 75 CRGs tested,

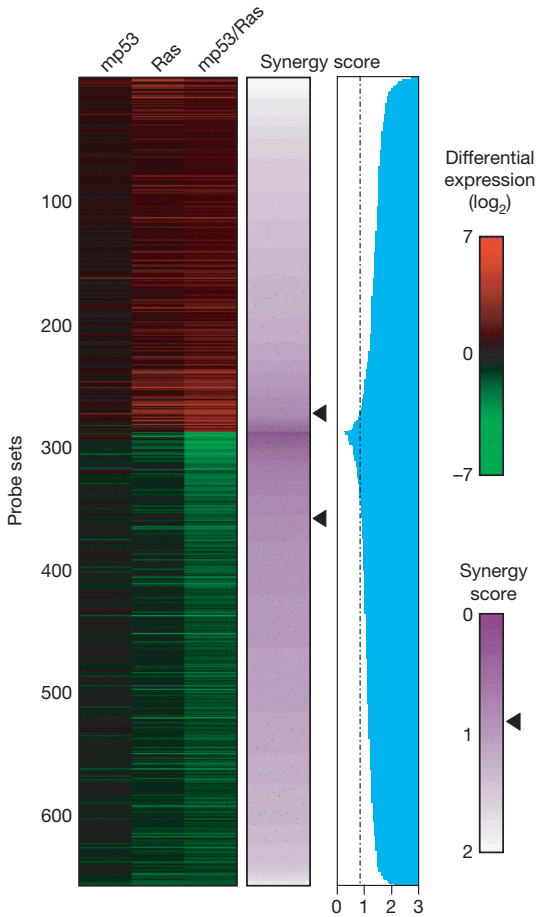


Figure 1 | Identification and characterization of cooperation response genes (CRGs). Raw expression values (log_2) of 538 differentially expressed genes (represented by 657 probe sets) for mp53, Ras and mp53/Ras cells, as compared to YAMC controls, are shown rank-ordered according to synergy score. Red and green indicate relative gene expression in the cells indicated versus YAMC cells; purple or blue indicate the synergy score for each gene plotted. A synergy score of 0.9 or less defines CRGs. The cutoff is indicated by arrowheads or the threshold line (stippled).

[1]Department of Biomedical Genetics, [2]Department of Biostatistics and Computational Biology, [3]James P. Wilmot Cancer Center, University of Rochester Medical Center, 601 Elmwood Avenue, Rochester, New York 14642, USA. [4]Department of Probability and Statistics, Charles University, Sokolovska 83, Praha-8, CZ-18675, Czech Republic.
*These authors contributed equally to this work
‡Deceased.

co-regulation is found in primary human colon cancer and our murine colon cancer cell model (Fig. 2c and Supplementary Fig. 3). Furthermore, altered expression of 29 CRGs has been reported in a variety of human cancer types, consistent with the direction of the change in gene expression observed in our experiments (Fig. 2c, Supplementary Table 1 and references therein). Thus, modulation of CRG expression has common features in malignant cell transformation of both murine and human cells.

The relevance of differentially expressed genes for malignant cell transformation was assessed by genetic perturbation of a series of 24 CRGs and 14 genes responding to $p53^{175H}$ and/or activated H-RasV12 in a non-cooperative manner (non-CRGs). Perturbed genes were chosen across a broad range of biological functions, levels of differential expression and synergy scores (Fig. 2, Supplementary Fig. 4 and Supplementary File 3). Gene perturbations were carried out in mp53/Ras cells with the goal of re-establishing mRNA expression of the manipulated genes to levels relatively close to those found in YAMC control cells, and to monitor subsequent tumour formation following subcutaneous injection of these cells into immunocompromised mice. Of the perturbed genes, 18 were upregulated and 20 were downregulated in mp53/Ras cells, relative to YAMC.

Reversal of the changes in CRG expression significantly reduced tumour formation by mp53/Ras cells in 14 out of 24 cases (Fig. 3a, left panel, Fig. 4a, c, Supplementary Fig. 5a and Supplementary Table 3), indicating a critical role in malignant transformation for a surprisingly large fraction of these genes. Perturbation of *Plac8*, *Jag2* and *HOXC13* gene expression had the strongest effects. We also combined perturbations of two CRGs, *Fas* and *Rprm*, that alone produced significant yet milder changes in tumour formation. Combination of these two CRGs yielded significantly increased efficacy in tumour

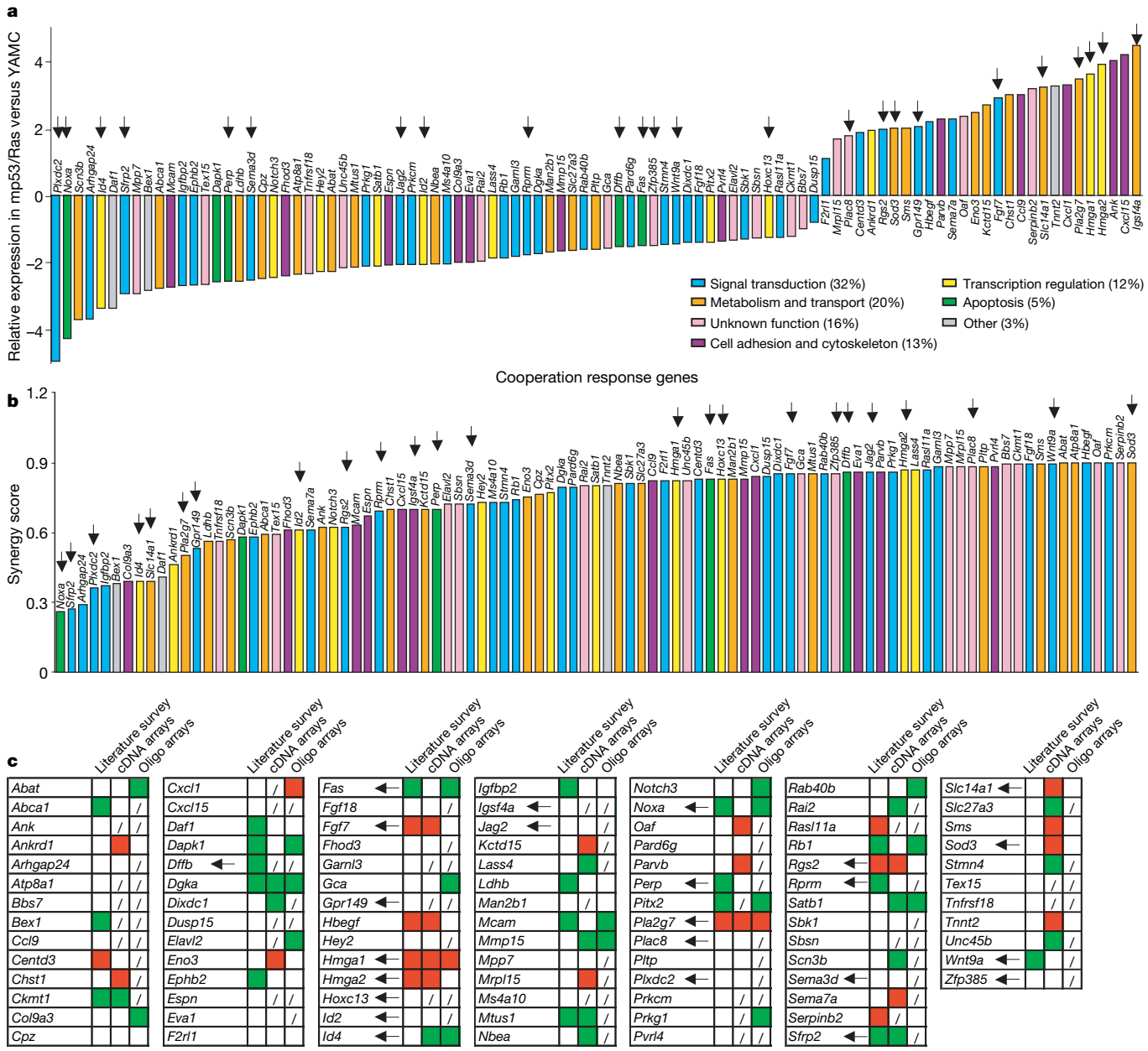


Figure 2 | Differential expression and synergy scores of CRGs in mp53/Ras cells and CRG co-regulation in human colon cancer. **a**, **b**, Bar graphs ranking CRG expression measured by microarray in mp53/Ras versus YAMC cells (**a**) and CRG synergy scores (**b**). Bars are colour-coded for gene-associated biological processes according to the Gene Ontology database. **c**, Table summarizing co-regulation of CRGs in mp53/Ras cells and human cancer on the basis of a literature survey for a variety of human cancers and two independent expression analyses of primary human colon cancers. Upregulation and downregulation of CRG expression compared to controls is indicated by red and green, respectively; lack of CRG representation on arrays is denoted by a solidus (/). Arrows indicate genes perturbed in this study.

inhibition as compared with the respective single perturbations (Fig. 4e, Supplementary Fig. 5b and Supplementary Table 4). Therefore, even genetic perturbations of CRGs with relatively smaller effects when examined on their own show evidence of being essential when analysed in combination.

In contrast to the multitude of CRG-related effects on tumour inhibition, out of the 14 non-CRG perturbations, only one showed a significant reduction in tumour formation of mp53/Ras cells (Fig. 3a, right panel, Supplementary Fig. 6 and Supplementary Table 5). Taken together, our data suggest that among the genes differentially expressed in cancer cells, malignant transformation strongly relies on the class of genes synergistically regulated by cooperating oncogenic mutations (Fig. 3b and Supplementary Fig. 7).

Genetic perturbation experiments were carried out using retrovirus-mediated re-expression of corresponding complementary DNAs (cDNAs) for downregulated genes (Supplementary Table 6) and short hairpin RNA (shRNA)-dependent stable knockdown using multiple independent targets for overexpressed genes (Supplementary Table 7). In addition, *Plac8* knockdown was functionally rescued by expression of shRNA-resistant *Plac8* (Fig. 4a), confirming the specificity of the *Plac8* loss-of-function experiments. The extent of all gene perturbations was assessed by qPCR (Supplementary Fig. 8). As expected, the genetic perturbations disrupt tumour formation downstream of the initiating oncogenic mutations. Expression of both mutant p53 and activated Ras proteins remained unaffected by all genetic manipulations that alter the formation of tumours (Supplementary Fig. 9). Moreover, gene perturbations distinguished tumour growth from *in vitro* cell proliferation, as they generally did not affect cell accumulation in tissue culture (Supplementary Fig. 10).

Perturbations of CRGs in human cancer cells (Fig. 4b, d, f, Supplementary Fig. 11 and Supplementary Tables 8, 9) had similarly strong tumour inhibitory effects to those in the genetically tractable murine mp53/Ras cells, as assessed by xenografts in nude mice. Perturbations of both upregulated and downregulated CRGs (that is, *Dffb*, *Fas*, *HOXC13*, *Jag2*, *Perp*, *PLAC8*, *Rprm*, *Zfp385* and *Fas/Rprm*) were performed in human DLD-1 and/or HT-29 colon cancer cell lines using retroviruses (Supplementary Fig. 12, Supplementary Tables 6 and 10) as described above. Similar to mp53/Ras cells, both human cancer cell lines have p53 mutations, whereas with K-Ras (DLD-1) and B-Raf (HT-29) mutations they express activated members of the Ras–Raf signalling pathway distinct from activated H-Ras in mp53/Ras cells. In addition, DLD-1 and HT-29 cells carry further oncogenic lesions such as APC and PIK3CA mutations, with HT-29 cells also having a mutation in *SMAD4* (see Supplementary Methods for references). The genetic perturbations had no effect on mutant Ras–Raf or p53 protein expression levels in both DLD-1 and HT-29 cells (Supplementary Fig. 13), indicating disruption of the cancer phenotype downstream of oncogenic mutations. Taken together, these experiments indicate the relevance of CRGs to cancer in a variety of backgrounds and genetic contexts.

The data described here indicate that the cooperative nature of malignant cell transformation depends, to a considerable degree, on a class of downstream effector genes regulated synergistically by multiple oncogenic mutations. We show that these CRGs contain a large fraction of genes (14 out of 24 tested) that are critical to the malignant phenotype, and that their perturbation—singly or in combination—can inhibit formation of tumours containing multiple oncogenic lesions, including p53 deficiency. In contrast, few of the genes differentially expressed in a non-synergistic manner (1 out of 14) significantly reduced tumour growth on perturbation. Synergistic behaviour found in gene expression data thus seems highly informative for identification of genes that are critically involved in malignant cell transformation (Fig. 3b), and provides a rational path to discovery of both cancer-cell-specific vulnerabilities and targets for intervention in cancer cells harbouring multiple mutations, including p53 loss-of-function.

CRGs represent a set of 95 annotated cellular genes, many of which have been associated with human cancer by virtue of altered gene

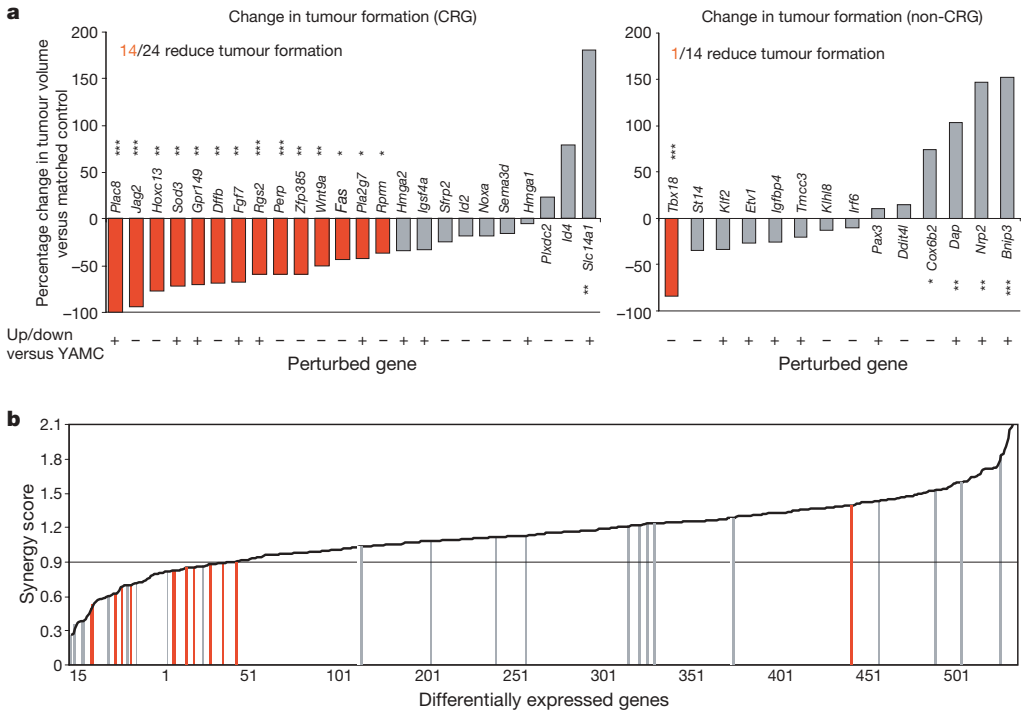


Figure 3 | Synergistic response of downstream genes to oncogenic mutations is a strong predictor for critical role in malignant transformation. a, Bar graphs indicating the percentage change in endpoint tumour volume after CRG and non-CRG perturbations in mp53/Ras cells (left and right panel, respectively). Perturbations significantly decreasing tumour size, as compared to matched controls, are shown in red. $***P < 0.001$, $**P < 0.01$, $*P < 0.05$; Wilcoxon signed-ranks and *t*-test. **b**, Distribution of gene perturbations over the set of genes differentially expressed in mp53/Ras cells, rank-ordered by synergy score. Bars (colour-coded as above) indicate perturbed genes; CRG cutoff synergy score (0.9) is indicated by a horizontal line.

expression (Fig. 2c and Supplementary Table 1). They are involved in the regulation of cell signalling, transcription, apoptosis and metabolism, and on the basis of our data represent key control points in many facets of cancer cell behaviour. We thus consider CRGs as critical nodes in gene networks underlying the malignant phenotype, providing an attractive rationale to explain why several features of cancer cells emerge simultaneously out of the interaction of a few genetic lesions[17].

Among CRGs and other differentially expressed effector genes we have also identified examples of genes that when perturbed produce significantly larger tumours (Fig. 3 and Supplementary Tables 3 and 5). This is consistent with the notion that oncogenic mutations can strongly induce antiproliferative cellular stress responses[18–21]. The existence of genes that restrict tumour formation while responding to oncogenic mutations supports the idea that the state of malignant transformation occurs as the result of a finely tuned balance between opposing signals generated by oncogenic mutations[15–17,20,22,23]. It is thus reasonable to speculate that tumour suppression via perturbation of CRGs as we have demonstrated may involve the disruption of this delicate balance. In fact, such targeted disruption downstream of oncogenic mutations may allow selective cancer cell deconstruction, yielding intervention strategies with high specificity for cancer cells.

For the 14 CRGs with tumour-inhibitory perturbations, a clear causal role in tumour formation downstream of oncogenic mutations has been shown here, to our knowledge, for the first time. Moreover, our data indicate that both gene extinctions (eight genes) and gene inductions (six genes) have important roles in this process. For example, we show that re-expression of the downregulated CRGs *Jag2* (a Notch ligand) or *HOXC13* (a homeobox transcription factor), as well as shRNA-dependent knockdown of *Plac8* gene expression, are each strongly tumour inhibitory in p53-defective murine and human cancer cells. Both Notch signalling[24] and HOXC13 (ref. 25) can have oncogenic functions in haematopoietic malignancies but are involved in promoting differentiation of epithelial cells[26,27], consistent with the tumour-inhibitory function of *Jag2* and *HOXC13* in the context of the solid-tumour models investigated here. *Plac8* is a little-investigated gene encoding a cysteine-rich highly conserved peptide, expressed in placenta, haematopoietic and epithelial cells, which is non-essential for mouse development[28]. *Plac8* can suppress p53 when overexpressed[29]; however, its essential role for tumour formation of p53-deficient cancer cells is unexpected. Among the eight downregulated CRGs is *Zfp385*, another gene of unknown function. Moreover, there are several pro-apoptotic and/or anti-proliferative genes such as *Perp*, *Rprm*, *Fas*, *Dffb* and *Wnt9a*, indicating that Ras activation and p53 deficiency cooperate to extinguish the expression of multiple growth-inhibitory genes, each of which contributes significantly to restricting tumour growth in the YAMC model when re-expressed. Out of these genes, *Perp*, *Rprm* and *Fas* have been previously identified as direct p53 targets, suggesting that their regulation by p53 is highly conditional on Ras activity (Supplementary Table 1 and references therein). Most of the upregulated CRGs contributing to tumour growth affect signal transduction. This includes *Fgf7*, *Rgs2*, *Gpr149*, an uncharacterized orphan seven-transmembrane receptor, and *Sod3*, which acts on signalling via modulation of metabolites[30]. Here we show a role in promoting tumour growth for all of these genes (including *Pla2g7*).

The efficacy of CRG perturbations performed in human colon cancer cells was comparable to that in the murine colon cell transformation model, suggesting dependence of the malignant state on a similar set of genes in both backgrounds. This is notable in light of the fact that these human cancer cells carry oncogenic mutations in genes in addition to *Ras* or *Raf* and *p53*, and suggests that CRGs may have a large involvement in the generation and maintenance of the cancer cell phenotype in a variety of contexts. CRGs may therefore provide a valuable source for identification by rational means of the much sought 'Achilles' heel' in human cancer.

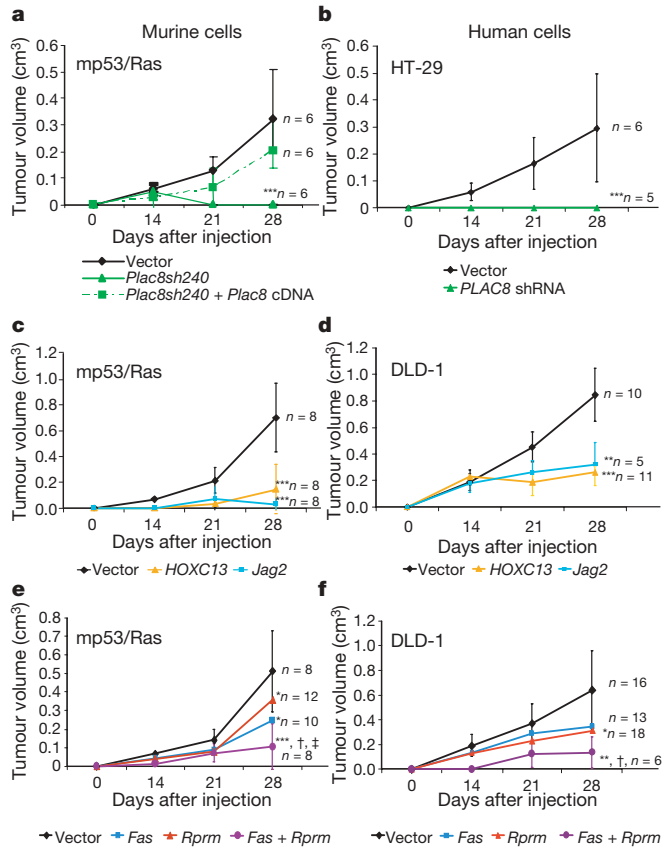


Figure 4 | CRG perturbations reduce tumour formation of both mp53/Ras and human cancer cells. a–f, Tumour volume was measured weekly for 4 weeks after injection into nude mice of murine (**a**, **c**, **e**) and human cancer cells (**b**, **d**, **f**) with indicated perturbations. Error bars indicate standard deviation at each time point. Number of injections (n) and significance levels as compared to matched controls are indicated; $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. Significance of tumour reduction on combined perturbation (*Fas* + *Rprm*) as compared to individual perturbations is indicated as follows: versus *Fas* ($\dagger P < 0.05$), versus *Rprm* ($\ddagger P < 0.05$).

METHODS SUMMARY

Cells. YAMC cells and derivation of cells with multiple oncogenic lesions[17] are described in Supplementary Information.

Microarray experiments, statistical analysis and CRG identification. Polysomal RNA was collected to obtain gene expression profiles reflective of protein synthesis rates. Expression values were obtained using the RMA procedure with background correction in Bioconductor (http://www.bioconductor.org). Differentially expressed genes were identified by the step-down Westfall–Young procedure in conjunction with the permutation N-test; the family-wise error rate was less than 0.01. Genes that respond synergistically to the combination of mutant p53 and activated Ras (CRGs) were selected by the following procedure. Let a represent mean expression value for a given gene in mp53 cells, b correspond to the mean expression value for the same gene in Ras cells, and d signify mean expression value for this gene in mp53/Ras cells. Then, the criterion defines CRGs as $\frac{a+b}{d} \leq 0.9$ for genes overexpressed in mp53/Ras cells and as $\frac{d}{a}+\frac{d}{b} \leq 0.9$ for genes underexpressed in mp53/Ras cells, as compared to controls. To assess robustness of synergy scores, jackknife sub-sampling was used to generate estimated P values for these scores. TaqMan low-density arrays (Applied Biosystems) were used to test gene expression differences observed by Affymetrix arrays independently.

Genetic perturbation of gene expression. cDNAs expressed via pBabe retroviral vectors or shRNA in pSuper-retro vectors were used to generate gene perturbations. These were tested by comparison of RNA expression levels in empty vector-infected cells and cells subjected to gene perturbation via SYBR Green qPCR with gene-specific primers.

Xenograft assays. Tumour formation was assessed by subcutaneous injection of cells into CD-1 nude mice (Crl:CD-1-Foxn1^{nu}, Charles River Laboratories). Tumour size was measured by caliper at 2, 3 and 4 weeks after injection.

Significance of difference in tumour size was calculated by means of the Wilcoxon signed-ranks test and the *t*-test, using directly matching vector control cells for each perturbation.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 September 2007; accepted 8 April 2008.
Published online 25 May 2008.

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
2. Yu, J. *et al.* Identification and classification of p53-regulated genes. *Proc. Natl Acad. Sci. USA* **96,** 14517–14522 (1999).
3. Zhao, R. *et al.* Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* **14,** 981–993 (2000).
4. Schulze, A., Lehmann, K., Jefferies, H. B., McMahon, M. & Downward, J. Analysis of the transcriptional program induced by Raf in epithelial cells. *Genes Dev.* **15,** 981–994 (2001).
5. Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genet.* **34,** 226–230 (2003).
6. Boiko, A. D. *et al.* A systematic search for downstream mediators of tumor suppressor function of p53 reveals a major role of BTG2 in suppression of Ras-induced transformation. *Genes Dev.* **20,** 236–252 (2006).
7. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408,** 307–310 (2000).
8. Vousden, K. H. & Lu, X. Live or let die: the cell's response to p53. *Nature Rev. Cancer* **2,** 594–604 (2002).
9. Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nature Rev. Cancer* **3,** 11–22 (2003).
10. Rodriguez-Viciana, P. *et al.* Cancer targets in the Ras pathway. *Cold Spring Harb. Symp. Quant. Biol.* **70,** 461–467 (2005).
11. Okada, F. *et al.* Impact of oncogenes in tumor angiogenesis: mutant K-ras up-regulation of vascular endothelial growth factor/vascular permeability factor is necessary, but not sufficient for tumorigenicity of human colorectal carcinoma cells. *Proc. Natl Acad. Sci. USA* **95,** 3609–3614 (1998).
12. Clark, E. A., Golub, T. R., Lander, E. S. & Hynes, R. O. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* **406,** 532–535 (2000).
13. Yang, J. *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117,** 927–939 (2004).
14. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436,** 518–524 (2005).
15. Sewing, A., Wiseman, B., Lloyd, A. C. & Land, H. High-intensity Raf signal causes cell cycle arrest mediated by $p21^{Cip1}$. *Mol. Cell. Biol.* **17,** 5588–5597 (1997).
16. Lloyd, A. C. *et al.* Cooperating oncogenes converge to regulate cyclin/cdk complexes. *Genes Dev.* **11,** 663–677 (1997).
17. Xia, M. & Land, H. Tumor suppressor p53 restricts Ras stimulation of RhoA and cancer cell motility. *Nature Struct. Mol. Biol.* **14,** 215–223 (2007).
18. Ridley, A. J., Paterson, H. F., Noble, M. & Land, H. Ras-mediated cell cycle arrest is altered by nuclear oncogenes to induce Schwann cell transformation. *EMBO J.* **7,** 1635–1645 (1988).
19. Hirakawa, T. & Ruley, H. E. Rescue of cells from *ras* oncogene-induced growth arrest by a second, complementing, oncogene. *Proc. Natl Acad. Sci. USA* **85,** 1519–1523 (1988).
20. Fanidi, A., Harrington, E. A. & Evan, G. I. Cooperative interaction between *c-myc* and *bcl-2* proto-oncogenes. *Nature* **359,** 554–556 (1992).
21. Denoyelle, C. *et al.* Anti-oncogenic role of the endoplasmic reticulum differentially activated by mutations in the MAPK pathway. *Nature Cell Biol.* **8,** 1053–1063 (2006).
22. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and $p16^{INK4a}$. *Cell* **88,** 593–602 (1997).
23. Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432,** 307–315 (2004).
24. Houde, C. *et al.* Overexpression of the NOTCH ligand JAG2 in malignant plasma cells from multiple myeloma patients and cell lines. *Blood* **104,** 3697–3704 (2004).
25. Panagopoulos, I. *et al.* Fusion of the *NUP98* gene and the homeobox gene *HOXC13* in acute myeloid leukemia with t(11;12)(p15;q13). *Genes Chromosom. Cancer* **36,** 107–112 (2003).
26. Nicolas, M. *et al.* Notch1 functions as a tumor suppressor in mouse skin. *Nature Genet.* **33,** 416–421 (2003).
27. Godwin, A. R. & Capecchi, M. R. *Hoxc13* mutant mice lack external hair. *Genes Dev.* **12,** 11–20 (1998).
28. Ledford, J. G., Kovarova, M. & Koller, B. H. Impaired host defense in mice lacking ONZIN. *J. Immunol.* **178,** 5132–5143 (2007).
29. Rogulski, K. *et al.* Onzin, a c-Myc-repressed target, promotes survival and transformation by modulating the Akt-Mdm2-p53 pathway. *Oncogene* **24,** 7524–7541 (2005).
30. Fattman, C. L., Schaefer, L. M. & Oury, T. D. Extracellular superoxide dismutase in biology and medicine. *Free Radic. Biol. Med.* **35,** 236–256 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Bohmann, C. Jordan and M. Noble for discussion; C. Brower, A. Cardillo, A. Petenkaya, L. Salamone, A. Brooks, Y. Xiao, S. Welle and A. Rosenberg for assistance with microarray data analysis; A. Burgess, R. Whitehead, J. Filmus and L. Milner for materials; and L. Maquat for sharing equipment. This work was supported in part by NIH grants CA90663, CA120317, GM075299 and a James P. Wilmot Cancer Center pilot grant. H.R.M. was supported in part by NIH T32 CA09363, P.S. by NIH K99 LM009477. This work is dedicated to A. Yakovlev.

Author Contributions H.L. conceived and directed the project. H.R.M., E.R.S., G.C., C.K. and B.S. designed and carried out experiments. S.-R.C and L.N. carried out experiments. P.S. consulted on and performed statistical analysis of microarray and tumour formation data. L.K. and A.Y. designed statistical methods to analyse microarray data. H.R.M. and H.L. wrote the paper.

Author Information Microarray data are deposited in the NCBI GEO database under accession number GSE9199. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.L. (land@urmc.rochester.edu).

METHODS

Cells. YAMC cells[31] and derivation of YAMC cells with multiple oncogenic lesions[17] are described in the Supplementary Information. Briefly, four polyclonal cell populations, control (Bleo/Neo), mp53 ($p53^{175H}$/Neo), Ras (Bleo/H-RasV12) and mp53/Ras ($p53^{175H}$/H-RasV12), were derived by retroviral infection of low-passage YAMC cells. Human colon cancer cells HT-29 were obtained from the ATCC; DLD-1 cells were provided by J. Filmus.

Microarray experiments. Polysomal RNA was extracted from YAMC, Bleo/Neo, mp53/Neo, Bleo/Ras and mp53/Ras cells to obtain gene expression profiles reflective of protein synthesis rates. RNA was collected from ten replicates per cell population grown in non-permissive conditions for 48 h, followed by 24 h in media with 0% FBS to maximize the contribution of oncogenic signalling to gene expression. RNA was collected when cells were sub-confluent and all cell populations were actively cycling. Cells were lysed in extraction buffer (50 mM MOPS, 15 mM MgCl, 150 mM NaCl, 0.5% Triton X-100 with 100 μg ml^{-1} cycloheximide, 1 mg ml^{-1} heparin, 200 U ribonuclease inhibitor (2 μl ml^{-1} of buffer) and 2 mM phenylmethyl sulphonyl fluoride). Supernatants were applied to 10–50% sucrose gradients, centrifuged at 36,000 r.p.m. for 2 h at 4 °C and fractions were collected using an ISCO gradient fractionator at 254 nm. Polysome containing fractions were pooled and RNA was purified using the RNeasy Mini Kit (Qiagen) following the standard protocol for animal cells, except that sucrose fractions were mixed with 3.5 volumes of buffer RLT before binding to the RNeasy column. RNA was on-column DNase digested as part of the RNeasy RNA extraction protocol.

Five micrograms of RNA was reverse transcribed and labelled using the mAMP kit (Ambion), with the 1× amplification protocol. The antisense RNA (cRNA) yield was fragmented and hybridization cocktails were prepared using the Affymetrix standard protocol for eukaryotic target hybridization. Targets were hybridized to Affymetrix Mouse Genome 430 2.0 Expression Arrays at 45 °C for 16 h, washed and stained using Affymetrix Fluidics protocol EukGE-WS2v4_450 in the Fluidics Station 450. Arrays were scanned with the Affymetrix GeneChip Scanner 3000.

Statistical analysis and CRG identification. Expression values from the 50 microarrays processed were obtained using the RMA procedure with background correction in Bioconductor (http://www.bioconductor.org). Differentially expressed genes were identified by the step-down Westfall–Young procedure[32] in conjunction with the permutation *N*-test[33]. The latter test is non-parametric and does not require log-expression levels to be normally distributed. The family-wise error rate was controlled at a level of 0.01. Gene expression values derived from mp53/Ras RNA samples were compared to those from two control cell populations, YAMC and Bleo/Neo cells; differentially expressed genes within the intersection of both comparisons were selected for further analysis (mp53/Ras versus YAMC, $P < 0.01$; mp53/Ras versus Bleo/Neo, $P < 0.01$). This selection process was executed in parallel using both raw and quantile normalized expression values, with the genes forming the union of both procedures being selected for further analysis. Expressed sequence tags and 'transcribed loci' were rejected from the set of genes thus selected.

Genes that responded synergistically to the combination of mutant p53 and activated Ras (that is, with a fold-change larger than the sum of fold-changes induced by mutant p53 and activated Ras individually) were termed CRGs. The following procedure was applied in parallel to mean values of raw and quantile normalized expression measurements, with the genes forming the union of both procedures being selected as CRGs for further analysis. Let a be the mean expression value for a given gene in mp53 cells, b represent the mean expression value for the same gene in Ras cells and d represent the mean expression value for this gene in mp53/Ras cells. Then, the selection criterion defines CRGs as $\frac{a+b}{d} \leq 0.9$ for genes overexpressed in mp53/Ras cells and as $\frac{d}{a} + \frac{d}{b} \leq 0.9$ for genes underexpressed in mp53/Ras cells, as compared to controls. Unlike a similar criterion based on the general isobol equation[34], this criterion has no rigorous theoretical justification; this formulation however, is heuristically appealing and served well for the purposes of our study. To assess robustness of synergy scores, jackknife sub-sampling was used to generate estimated *P*-values for these scores. Gene-associated biological processes for CRGs were assigned according to Gene Ontology database[35].

TaqMan low-density array qPCR. The TaqMan low-density array (Applied Biosystems) consists of TaqMan qPCR reactions targeting the cooperation response genes available (Supplementary Table 2) and control genes (18S ribosomal RNA, glyceraldehyde-3-phosphate dehydrogenase) in a microfluidic card. *Becn1* (Applied Biosystems probe set Mm00517174_ml) was used as a reference gene (see Supplementary Methods for further information). TaqMan Low-Density Arrays were used (four replicates per sample) to test gene-expression differences observed by Affymetrix arrays independently.

Expression of CRGs in human colon cancer. Co-regulation of CRGs in mp53/Ras cells and colon human cancer was assessed by comparing the *t*-statistics of CRG expression data reported here with those of two independent analyses of primary human colon cancers using cDNA or oligonucleotide arrays[36,37], respectively.

Genetic perturbation of gene expression. For stable gene re-expression of downregulated genes, cDNA for each gene was cloned into the pBabe retroviral vector, which was used to produce ecotropic or pseudotyped retrovirus for infection of mp53/Ras, HT-29 or DLD-1 cells. Cells were drug selected to derive polyclonal cell populations for xenograft assays.

For stable gene knockdown of upregulated genes, shRNAs targeting each gene were cloned into the pSuper-retro retroviral vector, which was used as pBabe vectors above. The specificity of the *Plac8* knockdown was independently confirmed by expression of *Plac8* cDNA rendered shRNA-resistant by the introduction of appropriate silent mutations. This shRNA-resistant cDNA was cloned into the pBabe-hygro retroviral vector and introduced into mp53/Ras cells harbouring Plac8sh240 shRNA.

The efficiency of gene perturbations was tested by the comparison of RNA expression levels in empty vector-infected mp53/Ras cells and cells subjected to gene perturbation via SYBR Green qPCR with gene-specific primers. Re-expression or knockdown was also compared with the respective levels of RNA expression in YAMC control cells.

31. Whitehead, R. H., VanEeden, P. E., Noble, M. D., Ataliotis, P. & Jat, P. S. Establishment of conditionally immortalized epithelial cell lines from both colon and small intestine of adult *H-2K*^{b}-tsA58 transgenic mice. *Proc. Natl Acad. Sci. USA* **90,** 587–591 (1993).
32. Westfall, P. H. & Young, S. S. Resampling-based multiple testing: examples and methods for P-value adjustment (Wiley, New York, 1993).
33. Klebanov, L., Gordon, A., Xiao, Y., Land, H. & Yakovlev, A. A permutation test motivated by microarray data analysis. *Comput. Stat. Data Anal.* **50,** 3619–3628 (2006).
34. Berenbaum, M. C. What is synergy? *Pharmacol. Rev.* **41,** 93–141 (1989).
35. The Gene Ontology Consortium.. Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
36. Saaf, A. M. *et al.* Parallels between global transcriptional programs of polarizing Caco-2 intestinal epithelial cells *in vitro* and gene expression programs in normal colon and colon cancer. *Mol. Biol. Cell* **18,** 4245–4260 (2007).
37. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98,** 15149–15154 (2001).

LETTERS

Deficiency in catechol-O-methyltransferase and 2-methoxyoestradiol is associated with pre-eclampsia

Keizo Kanasaki[1], Kristin Palmsten[1], Hikaru Sugimoto[1], Shakil Ahmad[2,3], Yuki Hamano[1], Liang Xie[1], Samuel Parry[4], Hellmut G. Augustin[5], Vincent H. Gattone Jr[6], Judah Folkman[7]‡, Jerome F. Strauss[8] & Raghu Kalluri[1,9,10]

Despite intense investigation, mechanisms that facilitate the emergence of the pre-eclampsia phenotype in women are still unknown. Placental hypoxia, hypertension, proteinuria and oedema are the principal clinical features of this disease. It is speculated that hypoxia-driven disruption of the angiogenic balance involving vascular endothelial growth factor (VEGF)/placenta-derived growth factor (PLGF) and soluble Fms-like tyrosine kinase-1 (sFLT-1, the soluble form of VEGF receptor 1) might contribute to some of the maternal symptoms of pre-eclampsia[1–5]. However, pre-eclampsia does not develop in all women with high sFLT-1 or low PLGF levels, and it also occurs in some women with low sFLT-1 and high PLGF levels[5,6]. Moreover, recent experiments strongly suggest that several soluble factors affecting the vasculature are probably elevated because of placental hypoxia in the pre-eclamptic women, indicating that upstream molecular defect(s) may contribute to pre-eclampsia. Here we show that pregnant mice deficient in catechol-O-methyltransferase (COMT) show a pre-eclampsia-like phenotype resulting from an absence of 2-methoxyoestradiol (2-ME), a natural metabolite of oestradiol that is elevated during the third trimester of normal human pregnancy. 2-ME ameliorates all pre-eclampsia-like features without toxicity in the *Comt*$^{-/-}$ pregnant mice and suppresses placental hypoxia, hypoxia-inducible factor-1α expression and sFLT-1 elevation. The levels of COMT and 2-ME are significantly lower in women with severe pre-eclampsia. Our studies identify a genetic mouse model for pre-eclampsia and suggest that 2-ME may have utility as a plasma and urine diagnostic marker for this disease, and may also serve as a therapeutic supplement to prevent or treat this disorder.

2-ME, a natural metabolite of oestradiol, is generated by COMT (Fig. 1a) in the placenta[7] and increases in concentration during pregnancy[8]. 2-ME destabilizes microtubules and inhibits hypoxia-inducible factor-1α (HIF-1α)[9–11], a transcription factor that senses tissue oxygen tension and regulates the expression of hypoxia-induced genes[12]. Placental COMT activity is suppressed in women with severe pre-eclampsia[13].

Comt$^{-/-}$ mice delivered preterm (19 ± 0 days after confirmation of a vaginal plug) in comparison with wild-type (*Comt*$^{+/+}$) mice (20.4 ± 0.12 days). *Comt*$^{-/-}$ mice also showed higher fetal wastage at day 17 of gestation in comparison with control wild-type mice (Fig. 1b). The total number of embryos was higher in the *Comt*$^{-/-}$ mice than in the wild-type control mice (Supplementary Fig. 1a). We speculate that this might be due to a compensatory mechanism resulting from augmented ovulation as a result of local ovarian catecholamines and catecholoestrogens[14].

The half-life of 2-ME is only about 20 min in rodents and significantly longer in humans[15,16]. We demonstrated that in the wild-type mice the concentration of 2-ME increased towards the end of pregnancy, with minimal levels of 2-ME observed in the pregnant *Comt*$^{-/-}$ mice (Fig. 1c). Subcutaneous administration of 10 ng of 2-ME in *Comt*$^{-/-}$ mice increased the concentration of 2-ME to levels similar to those observed in the control wild-type pregnant mice on day 17 (Fig. 1c). Circulating concentrations of steroids can vary significantly from mouse to mouse, even in identical inbred strains, as a result of rapid clearance, especially in pregnant mice with a variable number of embryos. Despite such limitations, we are able to show higher levels of 2-ME in the wild-type mice than in the *Comt*$^{-/-}$ mice.

Administration of 2-ME decreased fetal wastage in the *Comt*$^{-/-}$ mice (Fig. 1b and Supplementary Fig. 1a). To evaluate further the pregnancy status and embryonic development, we weighed placentae/decidua and embryos on day 17 of pregnancy. The placentae/decidua of *Comt*$^{-/-}$ mice weighed less than those of the wild-type mice, and weights were restored to normal by the administration of 2-ME (Fig. 1d). Embryos from control, *Comt*$^{-/-}$ and 2-ME-treated *Comt*$^{-/-}$ mice showed insignificant differences in their weights on day 17 of the pregnancy (Supplementary Fig. 1b). Moreover, normal pregnant mice treated with an inhibitor of COMT (Ro41-0960; 25 mg kg^{-1}) had significantly smaller placentae/decidua than the control mice (Fig. 1d). The embryo-to-placenta/decidua weight ratio was increased in the *Comt*$^{-/-}$ and inhibitor-treated wild-type mice, and 2-ME corrected this ratio in the *Comt*$^{-/-}$ mice (Supplementary Fig. 1c). These results suggest that COMT, by means of its production of 2-ME, is important for the appropriate development of the placenta and embryo. In this regard, the administration of oestradiol or 2-hydroxyoestradiol did not rescue the pregnant *Comt*$^{-/-}$ phenotype, indicating that the effect is specific to 2-ME (data not shown).

Placental insufficiency is speculated to contribute directly to the development of pre-eclampsia[17]. At the 14th day of pregnancy there were insignificant differences in the number of dead embryos between the control (0.83 ± 0.12) and *Comt*$^{-/-}$ (1.25 ± 0.15) mice. The placentae, as judged by the status of the labyrinth cells, spongiotrophoblasts and giant cells, appeared normal (Supplementary Fig. 2). *In situ* hybridization with placental-cell-specific markers revealed

[1]Division of Matrix Biology, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. [2]Departments of Reproductive and Vascular Biology Institute of Biomedical Research, The Medical School, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [3]Birmingham Women's Hospital, Edgbaston, Birmingham B15 2TG, UK. [4]Department of Obstetrics and Gynecology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104-6142, USA. [5]Joint Research Division Vascular Biology, Medical Faculty Mannheim, University of Heidelberg, and German Cancer Research Center Heidelberg, 69120 Heidelberg, Germany. [6]Department of Anatomy and Cell Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA. [7]Program in Vascular Biology, Department of Surgical Research, The Children's Hospital Boston, Boston, Massachusetts 02215, USA. [8]School of Medicine, Virginia Commonwealth University, Richmond, Virginia 23298, USA. [9]Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Boston, Massachusetts 02215, USA. [10]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02215, USA.
‡Deceased.

a normal expression and distribution of placental lactogen 1 (a marker for trophoblast giant cells), 4311 antigen (spongiotrophoblasts), Gcm-1 (syncytiotrophoblasts) and TFEB (chorionic trophoblasts) on day 17 in the pregnant *Comt*$^{-/-}$ and wild-type mice (Supplementary Fig. 3). Histological analysis revealed abundant eosin-positive deposition in the placentae/decidua of the *Comt*$^{-/-}$ mice (Fig. 1e, f). Furthermore, arteries in the decidua of *Comt*$^{-/-}$ mice contained hyaline-like deposits with foam cells in the wall and thrombosis in the lumen (32 thrombotic lesions were observed in 64 random placental/decidual sections from *Comt*$^{-/-}$ mice, in comparison with 2 of 54 for *Comt*$^{+/+}$ mice) (Fig. 1g, h). Vascular lesions in *Comt*$^{-/-}$ mice had non-specific IgM deposition in the vessel walls, in contrast with those of control wild-type mice (Supplementary Fig. 4a, b). Such vascular defects were not observed in any other organ in the pregnant *Comt*$^{-/-}$ mice (data not shown). The level of von Willebrand factor (vWF) was significantly diminished in the thrombotic vascular lesions of *Comt*$^{-/-}$ mice (Supplementary Fig. 4d, e), suggesting endothelial damage. The vascular abnormalities present in the *Comt*$^{-/-}$ mice were significantly diminished when 2-ME was administered (6 of 70 in *Comt*$^{-/-}$ mice administered with 2-ME) (Fig. 1i, j, and Supplementary Fig. 4c, f). The arteriopathy in the *Comt*$^{-/-}$ mice resembles decidual vascular lesions (acute atherosis) in women with pre-eclampsia[18]. Acute atherosis, together with thrombostic and fibrinoid changes, and infarction of the placenta are characteristic features of pre-eclamptic placenta/decidua. Our results strongly suggest that COMT/2-ME regulates utero-placental vascular homeostasis.

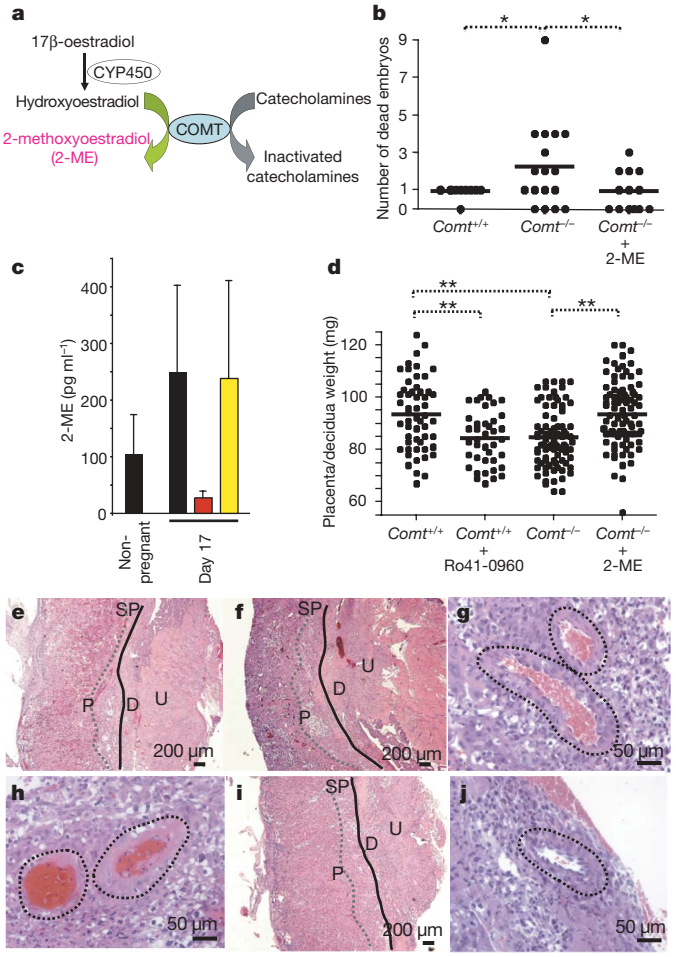


Figure 1 | Litter size and histological evaluation of placenta. **a**, Schematic illustration of 2-ME synthesis by COMT. **b**, Numbers of dead embryos at day 17 of pregnancy (wild-type mice, $n = 11$; *Comt*$^{-/-}$ mice, $n = 17$; *Comt*$^{-/-}$ mice with 2-ME, $n = 13$). Asterisk, $P < 0.05$. **c**, Plasma concentration of 2-ME in mice. Wild-type mice (black bars) showed an increase in 2-ME at 17 days of pregnancy, but *Comt*$^{-/-}$ mice (red bars) did not. *Comt*$^{-/-}$ mice + 2-ME (yellow bars) and wild-type mice had similar concentrations. Results show no significant difference (day 17 gestation wild-type mice versus *Comt*$^{-/-}$ mice, $P = 0.22$; day 17 gestation *Comt*$^{-/-}$ mice versus *Comt*$^{-/-}$ mice + 2-ME, $P = 0.22$). Non-pregnant wild-type mice, $n = 4$; day 17 gestation wild-type mice, $n = 5$; day 17 gestation *Comt*$^{-/-}$ mice, $n = 6$; day 17 gestation *Comt*$^{-/-}$ mice + 2-ME, $n = 6$. Results are shown as means and s.e.m. **d**, Weights of placentae. The numbers of placenta/decidua measured were as follows: wild-type mice, $n = 54$; *Comt*$^{+/+}$ mice + Ro41-0960, $n = 40$, *Comt*$^{-/-}$ mice, $n = 83$; *Comt*$^{-/-}$ mice + 2-ME, $n = 92$. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$. **e**, **f**, Placental/decidual histology in haematoxylin/eosin-stained sections from *Comt*$^{+/+}$ (**e**) and *Comt*$^{-/-}$ (**f**) mice. In **f**, eosin-positive depositions are found in placenta/decidua. The solid line indicates the border between placenta and decidua. **g**, **h**, Vascular lesions in the decidua in *Comt*$^{+/+}$ (**g**) and *Comt*$^{-/-}$ (**h**) mice. **i**, **j**, Samples from 2-ME-administered *Comt*$^{-/-}$ mice. Placenta (**i**) and vascular lesions (**j**) are shown. P, placenta; SP, spongiotrophoblast layer (a part of the placenta; between solid and dashed lines); D, decidua; U, uterus. The dashed line in **g**, **h** and **j** shows the vascular area in the decidua. Wild-type mice are designated as *Comt*$^{+/+}$ mice in the figure.

Hypertension associated with pre-eclampsia poses a serious risk to maternal health and to that of the newborn animal. In the wild-type mice, blood pressure decreased gradually throughout the pregnancy and a significant difference was observed between pregnant mice at day 17 and non-pregnant mice (Fig. 2a). This decrease in blood pressure was also observed during human pregnancy[19]. Non-pregnant wild-type and non-pregnant *Comt*$^{-/-}$ mice showed an insignificant difference in blood pressure (Fig. 2a). On day 10 of pregnancy, the blood pressure in *Comt*$^{-/-}$ mice began to fluctuate and increased significantly (Fig. 2a). On day 17 of pregnancy, the blood pressure of *Comt*$^{-/-}$ mice was significantly higher than that of wild-type pregnant mice and non-pregnant *Comt*$^{-/-}$ mice (Fig. 2a). The increased blood pressure observed in *Comt*$^{-/-}$ mice during the later stages of pregnancy returns to normal levels within 10 days after delivery, mimicking a similar trend observed in post-delivery pre-eclamptic women (Fig. 2a). Exogenous 2-ME prevented the increase in blood pressure in pregnant *Comt*$^{-/-}$ mice (Fig. 2a). An inhibitor of COMT (Ro41-0960; 25 mg kg^{-1}) administered to wild-type pregnant mice also elevates blood pressure during later stages of pregnancy in comparison with control wild-type pregnant mice (Fig. 2a). A *Comt*$^{-/-}$ female crossed with a wild-type male (placental genotype: *Comt*$^{+/-}$) did not show placental deficiency and hypertension (data not shown), suggesting that significant deficiency of COMT in the placenta is required for the development of a pre-eclampsia phenotype. Administration of 2-ME to non-pregnant wild-type mice and *Comt*$^{-/-}$ mice had an insignificant influence on the blood pressure status (Supplementary Fig. 5a, b).

Proteinuria is another clinical feature of pre-eclampsia. On day 17 of pregnancy, urinary albumin excretion in *Comt*$^{-/-}$ mice was significantly higher than that in pregnant wild-type control mice, non-pregnant control and *Comt*$^{-/-}$ mice (Fig. 2b). Wild-type mice treated with an inhibitor of COMT also excrete more urinary albumin than control mice in the later stages of pregnancy (Fig. 2b). Administration of 2-ME to pregnant *Comt*$^{-/-}$ mice prevents the occurrence of proteinuria (Fig. 2b).

By light-microscopic analysis, the kidney histologies of pregnant wild-type and pregnant *Comt*$^{-/-}$ mice were similar (Supplementary Fig. 6a, b). However, analysis by electron microscopy revealed kidney glomerular endothelial cell detachment, swelling and vacuolization (endotheliosis) on gestational day 17 of pregnant *Comt*$^{-/-}$ mice, and also in the COMT-inhibitor-treated pregnant wild-type mice (Fig. 2c–e). These ultrastructural glomerular lesions were not observed in pregnant 2-ME-treated *Comt*$^{-/-}$ mice (Fig. 2f). These results suggest that a deficiency in COMT/2-ME can result in the damage to the kidney glomerular structure, possibly through factors generated by placenta.

Other studies have suggested that placental hypoxia influences placental insufficiency in pre-eclampsia and may also contribute to the systemic disease due to defective angiogenesis[20]. In pregnant *Comt*$^{-/-}$ mice (day 17), both the placenta and decidua show signs of hypoxia (as determined by Hypoxyprobe incorporation) in comparison with those from control pregnant mice (Fig. 3a, b, d).

Hypoxia was seen predominantly in the placenta (spongiotrophoblast layer). (Fig. 3b). The markers of hypoxia were significantly diminished in $Comt^{-/-}$ mice receiving 2-ME (Fig. 3c, d). Our observations agree more closely with studies suggesting that hypoxia-induced HIF-1α protein accumulation in placentae is associated with shallow invasion of trophoblasts into the spiral arteries and uterine wall, probably resulting in vascular remodelling defects and further hypoxia[21]. In this regard, immunofluorescence and immunohistochemistry analysis revealed an increase in the accumulation of HIF-1α protein in placentae (spongiotrophoblast layer) of pregnant $Comt^{-/-}$ mice in comparison with pregnant control mice (Fig. 3e, f, and Supplementary Fig. 7a, b). HIF-1α accumulation was markedly decreased in placentae of 2-ME-treated pregnant $Comt^{-/-}$ mice (Fig. 3g and Supplementary Fig. 7c). Western blot analysis demonstrated that the level of HIF-1α protein in the nuclear fraction was higher in the placentae of pregnant $Comt^{-/-}$ mice than in placentae from the pregnant control mice (Fig. 3h). The increased accumulation of nuclear-translocated HIF-1α in the placentae of pregnant $Comt^{-/-}$ mice was prevented by 2-ME administration (Fig. 3h).

Hypoxia induces HIF-1α and also sFLT-1 in placental trophoblasts[21,22]. In this regard, increased sFLT-1 is speculated to be a significant factor in vascular endothelial damage in pre-eclampsia[2]. Plasma concentrations of sFLT-1 are reported to be significantly raised in women with pre-eclampsia[4,23]. In this study we showed that sFLT-1 is higher in all pregnant mice than in non-pregnant mice (Fig. 3i). In pregnant $Comt^{-/-}$ mice, the sFLT-1 concentration was found to be significantly higher than in pregnant control mice, starting at day 14 of pregnancy (Fig. 3i). The increase in sFLT-1 levels in the pregnant $Comt^{-/-}$ mice was prevented by 2-ME administration (Fig. 3i). Our experiments suggest that 2-ME may restore hypoxia-induced disruption of the angiogenic balance in the pregnant $Comt^{-/-}$ mice.

Plasma concentrations of catecholamines in pregnant wild-type and $Comt^{-/-}$ mice were also measured. An insignificant difference in catecholamines was detected between normal pregnant mice and pregnant $Comt^{-/-}$ mice (Supplementary Fig. 8).

Altered levels of vasodilators such as adrenomedullin and endothelial nitric oxide synthase (eNOS) are also reported in pre-eclampsia[24,25]. We therefore measured the placental expression of adrenomedullin and eNOS by *in situ* hybridization. Adrenomedullin expression revealed an unremarkable difference between all groups of mice (Supplementary Fig. 9a–c). eNOS expression in the placentae (spongiotrophoblast layer) was significantly decreased in the $Comt^{-/-}$ mice, and 2-ME administration restored eNOS expression in $Comt^{-/-}$ mice (Supplementary Fig. 9d–i).

Inflammatory mediators and natural killer (NK) cells were also evaluated. Whereas an insignificant difference in serum inflammatory T helper type 1 cytokines (tumour necrosis factor (TNF)-α and interferon (IFN)-γ) was observed between normal pregnant and $Comt^{-/-}$ pregnant mice (Supplementary Fig. 10), levels of decidual IFN-γ and numbers of NK cells were significantly higher in $Comt^{-/-}$ pregnant mice than in normal pregnant mice (Supplementary Figs 11 and 12).

In addition, as shown in Supplementary Fig. 13, we demonstrate in various *in vitro* experiments that 2-ME suppresses hypoxia-induced HIF-1α and sFLT-1 in HTR-8 cytotrophoblast-like cells.

Previous studies showed that 2-ME is present at about 3 ng ml^{-1} (about 10 nM) in the circulation of normal pregnant women during the third trimester[8]. The circulating levels of 2-ME were lower in women with pre-eclampsia than in normal pregnant women (all the samples are from gestational age 22–29 weeks) (Fig. 4a). We also found that COMT protein expression (membrane-bound and soluble) was significantly lower in the third-trimester pre-eclamptic placentae than in the gestational age-matched control placentae (Fig. 4b,

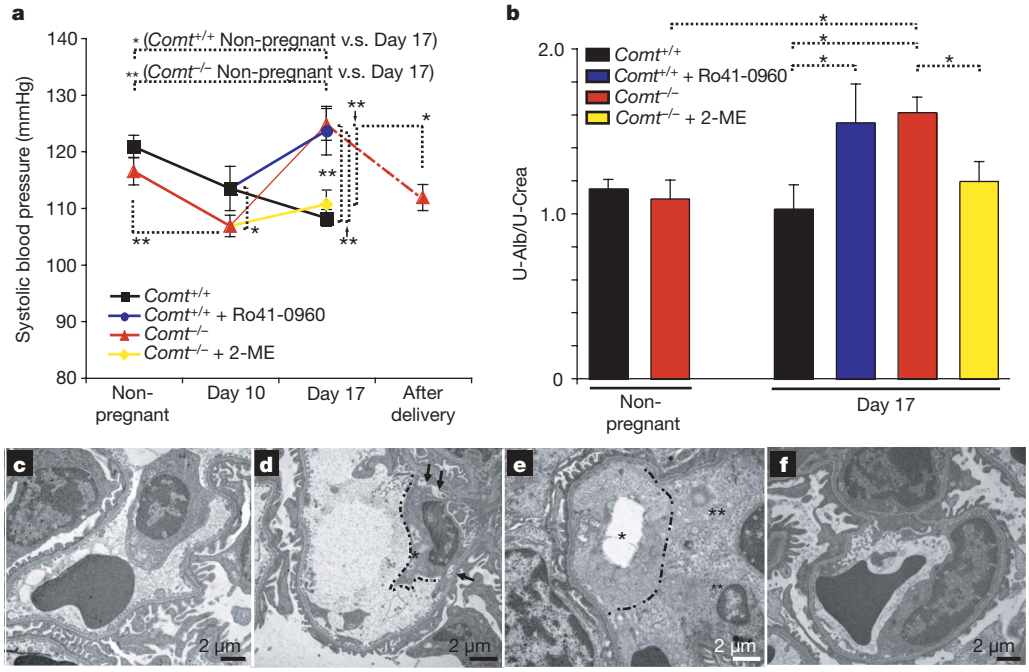


Figure 2 | Blood pressure and urine protein measurements. **a**, Systolic blood pressure was estimated in non-pregnant and pregnant mice on days 10 and 17 of gestation and 10 days after delivery. The following number of mice were evaluated: non-pregnant wild-type mice, $n = 11$; non-pregnant $Comt^{-/-}$ mice, $n = 15$; day 10 gestation wild-type pregnant mice, $n = 14$; day 10 gestation $Comt^{-/-}$ mice, $n = 22$; day 17 gestation wild-type mice, $n = 9$; day 17 gestation wild-type mice + Ro41-0960, $n = 6$; day 17 gestation $Comt^{-/-}$ mice, $n = 12$; day 17 gestation $Comt^{-/-}$ mice + 2-ME, $n = 10$; $Comt^{-/-}$ mice after delivery, $n = 9$. Results are shown as means ± s.e.m. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$. **b**, Urinary albumin and creatinine concentration was estimated with a colorimetric assay system. Urine albumin excretion was estimated as the quotient of urine albumin and urine creatinine. The following number of urine samples were evaluated: non-pregnant wild-type mice, $n = 7$; non-pregnant $Comt^{-/-}$ mice, $n = 7$; day 17 gestation wild-type mice, $n = 6$; wild-type mice + Ro41-0960, $n = 5$; day 17 gestation $Comt^{-/-}$ mice, $n = 11$; $Comt^{-/-}$ mice + 2-ME, $n = 9$. Results are shown as means and s.e.m. Asterisk, $P < 0.05$. **c**–**f**, Analysis of the sections from wild-type mice (**c**), wild-type mice + Ro41-0960 (**d**), $Comt^{-/-}$ mice (**e**) and $Comt^{-/-}$ mice + 2-ME (**f**) by electron microscopy. Glomerular endothelial cell detachment (**d**, arrow) with damage (**d**, dotted line), vacuolization (**d**, **e**, asterisk) and swelling (**e**, asterisk and two asterisks) are shown. Wild-type mice are designated as $Comt^{+/+}$ mice in the figure.

c, and Supplementary Table 1). HIF-1α is induced in placentae of pre-eclamptic women[26], suggesting that diminished COMT/2-ME levels could contribute to the elevation of placental HIF-1α. These studies highlight the need for a further comprehensive analysis of 2-ME levels in the urine and blood of women with pre-eclampsia.

COMT expression is reported to be robust in the human placenta and decidua basalis, as also observed in our study (Supplementary Fig. 14a–d)[7]. Our studies provide evidence for diminished levels of 2-ME in pre-eclamptic women. Although the precise role of placental/decidual COMT needs further investigation, our results provide us with a basis for proposing the following working model for the pathogenesis of pre-eclampsia. Disruption of COMT/2-ME, possibly due to variation in the *Comt* genotype, favours elevated levels of HIF-1α, leading to angiogenic dysfunction and placental insufficiency. HIF-1α elevation and vascular defects may lead to a shallow invasion of trophoblasts into the spiral arteries and uterine wall, resulting in vascular defects, hypoxia and inflammation[20,21]. Hypoxia and placental insufficiency may also lead to a deficiency in placenta-derived oestrogens and hydroxyoestradiols[27,28], which in turn may result in a further decrease in 2-ME level. Consequently, a vicious cycle is set in motion (Supplementary Fig. 15). Consistent with this model is a report that a functional *Comt* polymorphism causing low enzyme activity is associated with fetal growth restriction and abnormalities, which is a consequence of pre-eclampsia[29]. Our study highlights the potential use of 2-ME as both a diagnostic marker for pre-eclampsia and also as a therapeutic agent. (Additional discussion is included in Supplementary Information.)

METHODS SUMMARY

All sample collections from human were performed with informed consent from the patients, and sample collections were conducted under the approval of the South Birmingham Ethical Committee (UK), the University of Göttingen Ethical Committee and the University of Pennsylvania Ethical Committee. $Comt^{-/-}$ mice were developed by M. Karayiorgou and were provided under a Material Transfer Agreement[30]. At eight weeks of age, mating cages were set ($Comt^{+/+}$/$Comt^{+/+}$ or $Comt^{-/-}$/$Comt^{-/-}$). From day 10 of the pregnancy, mice were injected subcutaneously every day with 10 ng of 2-ME ($Comt^{-/-}$ mice) or 25 mg kg^{-1} Ro41-0960 (wild-type mice) or with placebo (olive oil). For non-invasive monitoring of blood pressure, mice were trained for at least five days before measurement of blood pressure. Mouse studies followed the Beth Israel Deaconess Medical Center (BIDMC) Institutional Animal Care and Use guidelines. Urinary albumin and creatinine measurement were performed as previously described. Measurement of the concentration of plasma 2-ME was performed by PPD Development. Tissue hypoxia was evaluated with a Hypoxyprobe-1 kit (Chemicon International) and detected by immunohistochemistry (Histomouse Max Kit; Zymed Laboratories Inc.). Tissue labelling and western blotting for HIF-1α, COMT, eNOS and vWF were performed with standard methods. IFN-γ measurement in the tissue was performed by immunohistochemistry. NK cell immunolabelling was performed as described previously. Plasma sFLT-1 and catecholamine concentrations and serum

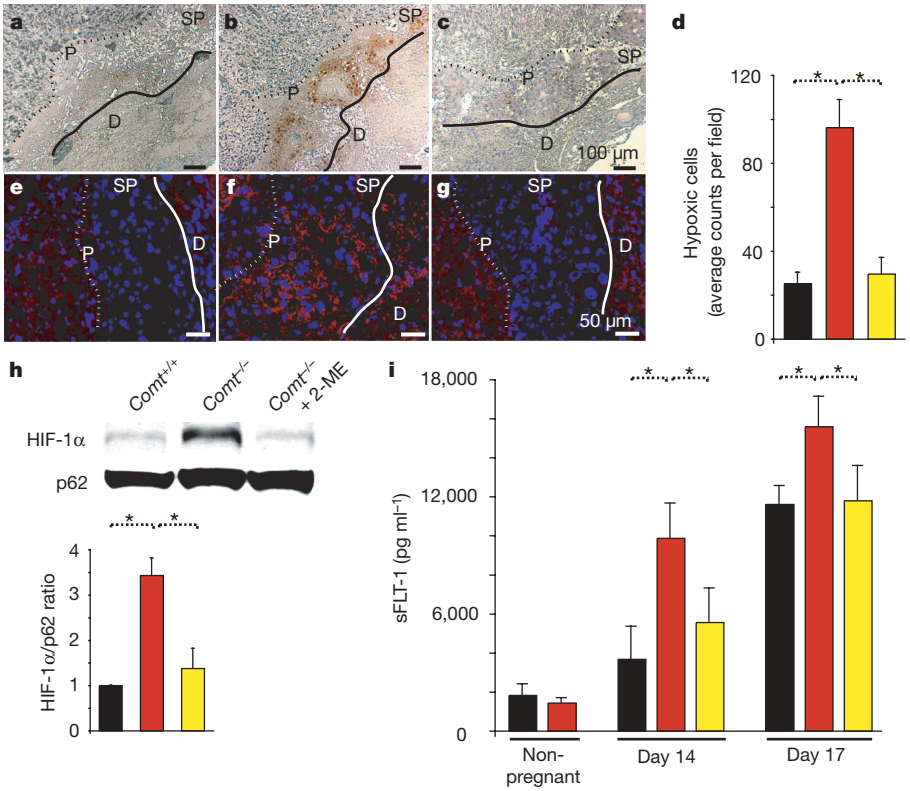


Figure 3 | Role of 2-ME in the placenta. **a–c**, Hypoxyprobe in placenta/decidua samples from wild-type mice (**a**), $Comt^{-/-}$ mice (**b**) and $Comt^{-/-}$ mice + 2-ME (**c**) was detected by immunohistochemistry. $Comt^{-/-}$ mice show hypoxia dominantly in the spongiotrophoblast layer of placentae. **d**, Quantitative measurement of hypoxic cells. Randomly selected 10 placenta/decidua were scanned at a magnification of ×100. The Hypoxyprobe-positive cells (brown cells) were counted in each field and an average was taken over 10 fields. Black bars, wild-type mice; red bars, $Comt^{-/-}$ mice; yellow bars, $Comt^{-/-}$ mice + 2-ME (the same colour scheme is used in **h** and **i**). Results are shown as means and s.e.m. Asterisk, $P < 0.05$. **e–g**, HIF-1α protein expression in the placenta from wild-type mice (**e**), $Comt^{-/-}$ mice (**f**) and $Comt^{-/-}$ mice + 2-ME (**g**) was evaluated by immunofluorescence microscopy. Merged images are shown here (HIF-1α plus nuclear 4,6-diamidino-2-phenylindole). the solid line indicates the border between placenta (P) and decidua (D). SP, Spongiotrophoblast layer (a part of the placenta; between solid and dashed line). Scale bars, 100 μm (**a**, **b**); 50 μm (**e**, **f**). **h**, Western blot analysis for HIF-1α using nuclear protein fraction from placenta. Nucleoporin p62 protein levels are shown, to indicate equal loading of nuclear proteins. Representative results from three independent experiments are shown. The lower panel shows a densitometric analysis of HIF-1α protein expression normalized to p62 protein. The results are shown as the relative expression against $Comt^{+/+}$ in each set of western blots. Results are shown as means and s.e.m. ($n = 3$). Asterisk, $P < 0.05$. **i**, ELISA for sFLT-1 analysis. Non-pregnant wild-type, $n = 6$; non-pregnant $Comt^{-/-}$, $n = 7$; day 14 gestation wild-type mice, $n = 6$; day 14 gestation $Comt^{-/-}$ mice, $n = 8$; day 14 gestation $Comt^{-/-}$ mice + 2-ME, $n = 6$; day 17 gestation wild-type mice, $n = 8$; day 17 gestation $Comt^{-/-}$ mice, $n = 12$; day 17 gestation $Comt^{-/-}$ mice + 2-ME, $n = 13$. Results are shown as means and s.e.m. Asterisk, $P < 0.05$. Wild-type mice are designated as $Comt^{+/+}$ mice in the figure.

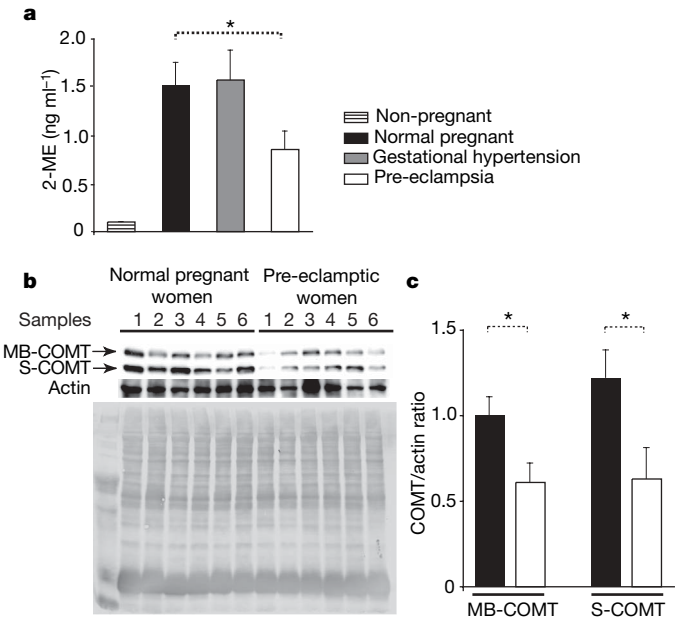


Figure 4 | COMT and 2-ME measurements during human pregnancy. **a**, Plasma concentration of 2-ME in non-pregnant women ($n = 2$), normal pregnant women ($n = 13$), women with gestational hypertension ($n = 9$) and pre-eclamptic women ($n = 8$). All samples were collected at 22–29 weeks of gestation. Results are shown as means and s.e.m. Asterisk, $P < 0.05$. **b**, Western blot analysis of COMT using protein lysates (10 μg) from human placenta (normal pregnant women, $n = 6$; pre-eclamptic women, $n = 6$). MB, membrane-bound; S, soluble. Actin protein levels are shown as a loading control. The lower panel shows Coomassie blue labelling of the membrane. **c**, Quantification (densitometric scan analysis) of the protein expression of COMT normalized by actin protein level in the placenta. The results are shown as expression relative to the mean expression of MB-COMT in normal pregnant women. Open bars, pre-eclamptic women; filled bars, normal pregnant women. Results are shown as means and s.e.m. Asterisk, $P < 0.05$.

cytokine concentrations were measured with an enzyme-linked immunosorbent assay (ELISA) system. RT–PCR was performed with standard methods. Electron microscopy was performed as described previously. *In situ* hybridization was performed with a DIG High Prime DNA Labelling and Detection system (Roche). *In vitro* experiments were performed with HTR-8 cells.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 December 2007; accepted 31 March 2008.
Published online 11 May 2008.

1. Vuorela, P. *et al.* Amniotic fluid–soluble vascular endothelial growth factor receptor-1 in preeclampsia. *Obstet. Gynecol.* **95,** 353–357 (2000).
2. Sugimoto, H. *et al.* Neutralization of circulating vascular endothelial growth factor (VEGF) by anti-VEGF antibodies and soluble VEGF receptor 1 (sFlt-1) induces proteinuria. *J. Biol. Chem.* **278,** 12605–12608 (2003).
3. Krauss, T., Pauer, H. U. & Augustin, H. G. Prospective analysis of placenta growth factor (PlGF) concentrations in the plasma of women with normal pregnancy and pregnancies complicated by preeclampsia. *Hypertens. Pregnancy* **23,** 101–111 (2004).
4. Maynard, S. E. *et al.* Excess placental soluble fms-like tyrosine kinase 1 (sFlt1) may contribute to endothelial dysfunction, hypertension, and proteinuria in preeclampsia. *J. Clin. Invest.* **111,** 649–658 (2003).
5. Levine, R. J. *et al.* Circulating angiogenic factors and the risk of preeclampsia. *N. Engl. J. Med.* **350,** 672–683 (2004).
6. Solomon, C. G. & Seely, E. W. Preeclampsia—searching for the cause. *N. Engl. J. Med.* **350,** 641–642 (2004).
7. Casey, M. L. & MacDonald, P. C. Characterization of catechol-O-methyltransferase activity in human uterine decidua vera tissue. *Am. J. Obstet. Gynecol.* **145,** 453–457 (1983).
8. Berg, D., Sonsalla, R. & Kuss, E. Concentrations of 2-methoxyoestrogens in human serum measured by a heterologous immunoassay with an ^{125}I-labelled ligand. *Acta Endocrinol. (Copenh.)* **103,** 282–288 (1983).
9. Fotsis, T. *et al.* The endogenous oestrogen metabolite 2-methoxyoestradiol inhibits angiogenesis and suppresses tumour growth. *Nature* **368,** 237–239 (1994).
10. D'Amato, R. J., Lin, C. M., Flynn, E., Folkman, J. & Hamel, E. 2-Methoxyestradiol, an endogenous mammalian metabolite, inhibits tubulin polymerization by interacting at the colchicine site. *Proc. Natl Acad. Sci. USA* **91,** 3964–3968 (1994).
11. Mabjeesh, N. J. *et al.* 2ME2 inhibits tumor growth and angiogenesis by disrupting microtubules and dysregulating HIF. *Cancer Cell* **3,** 363–375 (2003).
12. Semenza, G. L. Hypoxia-inducible factor 1: master regulator of O_2 homeostasis. *Curr. Opin. Genet. Dev.* **8,** 588–594 (1998).
13. Barnea, E. R., MacLusky, N. J., DeCherney, A. H. & Naftolin, F. Catechol-O-methyl transferase activity in the human term placenta. *Am. J. Perinatol.* **5,** 121–127 (1988).
14. Senthilkumaran, B. & Joy, K. P. Periovulatory changes in catfish ovarian oestradiol-17β, oestrogen-2-hydroxylase and catechol-O-methyltransferase during GnRH analogue-induced ovulation and *in vitro* induction of oocyte maturation by catecholoestrogens. *J. Endocrinol.* **168,** 239–247 (2001).
15. Ireson, C. R. *et al.* Pharmacokinetics and efficacy of 2-methoxyoestradiol and 2-methoxyoestradiol-bis-sulphamate *in vivo* in rodents. *Br. J. Cancer* **90,** 932–937 (2004).
16. Zacharia, L. C. *et al.* 2-hydroxyestradiol is a prodrug of 2-methoxyestradiol. *J. Pharmacol. Exp. Ther.* **309,** 1093–1097 (2004).
17. Redman, C. W. & Sargent, I. L. Latest advances in understanding preeclampsia. *Science* **308,** 1592–1594 (2005).
18. Labarrere, C. A. Acute atherosis. A histopathological hallmark of immune aggression? *Placenta* **9,** 95–108 (1988).
19. MacGillivray, I., Rose, G. A. & Rowe, B. Blood pressure survey in pregnancy. *Clin. Sci.* **37,** 395–407 (1969).
20. Genbacev, O., Joslin, R., Damsky, C. H., Polliotti, B. M. & Fisher, S. J. Hypoxia alters early gestation human cytotrophoblast differentiation/invasion *in vitro* and models the placental defects that occur in preeclampsia. *J. Clin. Invest.* **97,** 540–550 (1996).
21. Caniggia, I. *et al.* Hypoxia-inducible factor-1 mediates the biological effects of oxygen on human trophoblast differentiation through TGFβ$_3$. *J. Clin. Invest.* **105,** 577–587 (2000).
22. Nagamatsu, T. *et al.* Cytotrophoblasts up-regulate soluble fms-like tyrosine kinase-1 expression under reduced oxygen: an implication for the placental vascular development and the pathophysiology of preeclampsia. *Endocrinology* **145,** 4838–4845 (2004).
23. Ahmad, S. & Ahmed, A. Elevated placental soluble vascular endothelial growth factor receptor-1 inhibits angiogenesis in preeclampsia. *Circ. Res.* **95,** 884–891 (2004).
24. Myatt, L., Eis, A. L., Brockman, D. E., Greer, I. A. & Lyall, F. Endothelial nitric oxide synthase in placental villous tissue from normal, pre-eclamptic and intrauterine growth restricted pregnancies. *Hum. Reprod.* **12,** 167–172 (1997).
25. Hata, T., Miyazaki, K. & Matsui, K. Decreased circulating adrenomedullin in pre-eclampsia. *Lancet* **350,** 1600–1605 (1997).
26. Rajakumar, A., Brandon, H. M., Daftary, A., Ness, R. & Conrad, K. P. Evidence for the functional activity of hypoxia-inducible transcription factors overexpressed in preeclamptic placentae. *Placenta* **25,** 763–769 (2004).
27. Takanashi, K., Honma, T., Kashiwagi, T., Honjo, H. & Yoshizawa, I. Detection and measurement of urinary 2-hydroxyestradiol 17-sulfate, a potential placental antioxidant during pregnancy. *Clin. Chem.* **46,** 373–378 (2000).
28. Rosing, U. & Carlstrom, K. Serum levels of unconjugated and total oestrogens and dehydroepiandrosterone, progesterone and urinary oestriol excretion in pre-eclampsia. *Gynecol. Obstet. Invest.* **18,** 199–205 (1984).
29. Sata, F. *et al.* Functional maternal catechol-O-methyltransferase polymorphism and fetal growth restriction. *Pharmacogenet. Genomics* **16,** 775–781 (2006).
30. Gogos, J. A. *et al.* Catechol-O-methyltransferase-deficient mice exhibit sexually dimorphic changes in catecholamine levels and behavior. *Proc. Natl Acad. Sci. USA* **95,** 9991–9996 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We dedicate this manuscript to Judah Folkman for his inspiration and guidance. We thank P. T. Männistö for providing us with information about the *Comt*$^{-/-}$ mice. Beth Israel Deaconess Medical Center has licensed technologies associated with 2-ME/COMT and preeclampsia to Cynthus, Inc. This work was supported primarily by the BIDMC Department of Medicine research funds to the Division of Matrix Biology, and partly supported by National Institutes of Health grants DK 55001, DK 62987, DK 13193 and DK 61688. S.A. was supported by grants from the British Heart Foundation and the Medical Research Council. K.K. was partly supported by Foreign Study Grants from the Kanae Foundation for the Promotion of Medical Science in Japan.

Author Contributions K.K. performed all the experiments, analysed the data and participated in manuscript writing. K.P., H.S., Y.H. and L.X. performed some of the experiments. V.H.G. performed the EM analysis. S.P., S.A. and H.G.A. provided samples. J.F.S. provided samples and contributed to manuscript writing. J.F. made intellectual contributions. R.K. conceived the project, designed the experimental approach, made intellectual contributions and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.K. (rkalluri@bidmc.harvard.edu).

METHODS

Reagents. 2-ME, COMT inhibitor (Ro41-0960), MAO-A inhibitor (Clorgyline; *N*-methyl-*N*-propargyl-3-(2,4-dichlorophenoxy)propylamine hydrochloride), 17β-oestradiol and fluorescein isothiocyanate (FITC)-conjugated anti-mouse IgM (F-9259) were obtained from Sigma-Aldrich. Anti-HIF-1α polyclonal (for western blot analysis and immunohistochemistry; NB 100-449) and monoclonal (for immunofluorescence; NB 100-131) antibodies were purchased from Novus Biologicals. Anti-COMT polyclonal antibody and anti-α-tubulin antibody were purchased from Chemicon International. Anti-eNOS polyclonal antibody (ab5589) for western blot analysis was purchased from Abcam. Anti-vWF polyclonal antibody (A0082) was purchased from Dako. ELISA kit for sFLT-1, TNF-α, IFN-γ and goat polyclonal antibody against IFN-γ was purchased from R&D Systems Inc. and the ELISA kits for catecholamines were from IBL. FITC-conjugated anti-goat, rhodamine-conjugated anti-hamster, rhodamine-conjugated anti-mouse and FITC-conjugated anti-rabbit antibodies were obtained from Jackson Immunoresearch.

Tissue collection. Human placental tissues were obtained from uncomplicated term pregnancies delivered by elective caesarean section for breech presentation or a recurring indication in otherwise uncomplicated pregnancies, as described previously[23]. In addition, similar gestationally matched placental tissues were collected by elective caesarean section from pregnancies complicated by pre-eclampsia. For the placenta collection, pre-eclampsia was defined as a blood pressure of more than 140/90 mmHg on at least two consecutive measurements and proteinuria of at least 300 mg per 24 h. Informed consent was obtained from the patients and the study had the approval of the South Birmingham Ethical Committee.

Animals. Successful mating in *Comt*$^{-/-}$ mice was determined by the appearance of a vaginal plug. Midday of the day when the vaginal plug was observed was taken to be 12 h after fertilization, embryonic day 0.5 (E0.5). The mice with a confirmed vaginal plug were placed in a different cage (three or four mice to a cage) until being killed. From day 10 of the pregnancy, mice were injected daily with 10 ng of 2-ME (*Comt*$^{-/-}$ mice) or 25 mg kg^{-1} Ro41-0960 (wild-type mice) or placebo (olive oil), subcutaneously. Mouse studies followed the BIDMC Institutional Animal Care and Use Guidelines.

Measurements of 2-ME. The measurement of plasma 2-ME was performed by PPD Development. To determine concentrations of 2-ME, 100 µl of human or 25 µl mouse plasma samples were fortified with 50 µl of internal standard working solution. Methanol (10% in water) and hydrolysis buffer were added to the samples, and then the sample was vortex-mixed. After extraction, solvent was added, samples were vortex-mixed again and centrifuged. The aqueous layer was frozen and the organic layer was transferred to a clean tube. The extract was evaporated and the remaining residue was reconstituted with 200 µl of extraction buffer (0.1% formic acid, 16% acetonitrile, 4% methanol). A 50-µl volume of the final extract was injected and analysed by HPLC with MS/MS detection. The assay details are: recovery rate of 88.6% ($n = 9$), inter–assay precision and accuracy ($n = 24$), coefficient of variation <8.00%, difference from theory range −4.16 to 8.38% ($n = 6$ on four separate days), coefficient of variation <13.3%, difference from theory range −5.75 to 9.78%, and concentration range 1.00–1,000 ng ml^{-1}.

Informed consent was obtained from the patients and the study had the approval of the University of Göttingen Ethical Committee and also the University of Pennsylvania Ethical Committee. All samples used for circulating 2-ME measurements were from normal pregnant women, women with gestational hypertension or women with pre-eclampsia. All samples were obtained between 22 and 29 weeks of gestation.

Blood pressure monitoring. Mice were trained at least five days before the evaluation of blood pressure. The blood pressure of conscious mice was measured by a programmable tail-cuff sphygmomanometer (SC-1000; Hatteras Instruments). Blood pressure was measured in non-pregnant mice and pregnant mice at day 10 and 17 after plug confirmation, and also 10 days after delivery.

Urinary albumin and creatinine measurement. Albumin and creatinine measurements in the urine were estimated with a colorimetric assay in accordance with the manufacturer's recommendations (Sigma-Aldrich). Urine albumin excretion was estimated as the quotient of urine albumin and urine creatinine[31].

Immunofluorescence. Frozen sections were fixed in 100% acetone at −20 °C for 10 min. After blocking, sections were incubated with primary antibodies against HIF-1α (dilution 1:200) or COMT (1:400) for 1 h at room temperature (26 °C) and subsequently labelled with secondary antibodies.

Immunohistochemistry for COMT, vWF and HIF-1α. Deparaffinized (2 min in xylene, four times; 1 min in 100% ethanol, twice; 30 s in 95% ethanol, 45 s in 70% ethanol, 1 min in distilled water) placental/decidual sections were used for COMT labelling. Immunohistochemistry was performed with a Vectastain ABC Kit using DAB peroxidase substrate reagent (Vector Laboratories, Inc.). The primary antibody was used at the following dilutions: COMT, 1:200; vWF, 1:100; HIF-1α, 1:100. For vWF and HIF-1α staining, deparaffinized sections were treated with prewarmed proteinase K (20 µg ml^{-1}) incubation for 15 min before blocking.

Immunohistochemistry for IFN-γ. Frozen placental/decidual sections were used for the detection of IFN-γ with a Vectastain ABC Kit and DAB peroxidase substrate reagent. The primary antibody was used at a dilution of 1:100.

Detection of NK cells in placenta/decidua. Frozen placental/decidual sections were fixed with acetone (−20 °C) for 10 min. Fixed sections were washed once with PBS and then blocked with 2% BSA in PBS. Blocked sections were incubated with goat anti-NKp46 (R&D System) and hamster anti-CD3 (eBioscience) antibody (both at 1:50 dilution) in primary antibody dilution buffer (1% BSA, 0.1% porcine skin gelatin, 0.05% sodium azide and 0.01 M PBS pH 7.2) overnight at 4 °C. Next day, sections were washed three times with PBS, then incubated with FITC-conjugated anti-goat IgG and rhodamine-conjugated anti-hamster IgG secondary antibody (1:200 dilution) for 1 h at room temperature. Subsequently the sections were washed five times with PBS and once with distilled water. The sections were covered with mounting medium (Vector Laboratories). Immunofluorescence was detected by fluorescence microscopy. NK cells were identified as NKp46-positive and CD3-negative[32].

Electron microscopy. Kidney tissues were fixed in 0.1 M cacodylate with 2% glutaraldehyde. Electron microscopy was performed as described previously[31].

Tissue hypoxia evaluation. Tissue hypoxia was also evaluated with a Hypoxyprobe-1 kit (Chemicon International). Mice were injected intraperitoneally with hypoxic probe (60 mg kg^{-1} body weight), 3 h before being killed. Injected hypoxic probe in deparaffinized sections were detected by immunohistochemistry (Histomouse Max Kit; ZYMED Laboratories Inc.). Sections were counterstained with haematoxylin.

Hypoxic chamber. Cells were exposed to hypoxia (2% O_2) using a gas mixture (95% N_2/5% CO_2) in an air chamber for 16 h with or without co-incubation of test reagents.

Nuclear protein fraction from the placenta. The maternal component of placentae (that is, decidua and uterus) was stripped away and the conceptus-derived part of the placentae was used for nuclear extraction, as described elsewhere[33]. Placental samples were homogenized and lysed with hypotonic buffer (500 mM HEPES-KOH pH 7.5, 1% Nonidet P40 and protease inhibitor cocktail (Roche)), and the lysates were centrifuged at 3,000*g* for 10 min. The pellets were resuspended in high-salt buffer (hypotonic buffer with 500 mM NaCl and 25% glycerol), rotated for 30 min at 4 °C, and centrifuged at 17,000*g* for 30 min. The supernatant was used for western blot analysis. Equal amounts of nuclear proteins were determined by estimation of nucleoporin p62 (BD Transduction Laboratories).

Western blotting. Protein lysates (nuclear fraction for HIF-1α or purified protein with lysis buffer (50 mM Tris-HCl pH 7.5, 0.15 M NaCl, 0.1% SDS, 1% Triton X-100, 1% deoxycholate, with protease inhibitor)) were denatured with SDS-sample buffer in boiling water for 5 min. After centrifugation at 17,000*g* for 10 min at 4 °C, the supernatant was separated on 8% or 12% SDS-polyacrylamide gels, blotted onto poly(vinylidene difluoride) membranes (Immobilon) by a semidry method. After blocking with TBS-T (Tris-buffered saline containing 0.05% Tween 20) containing 5% non-fat dried milk, the membranes were incubated overnight at 4 °C with 1:500 diluted anti-HIF-1α, 1:2,000 anti-COMT and 1:1,000 anti-eNOS polyclonal antibody in TBST containing 5% BSA. The membranes were washed three times and incubated with 1:10,000 diluted horseradish-peroxidase-conjugated secondary antibody (Promega) at room temperature for 1 h. The immunoreactive bands were detected with an enhanced chemiluminescence (ECL) detection system (Pierce Biotechnology).

RT–PCR. RT–PCR was performed with SuperScript II (Invitrogen). RNA (4 µg) from placenta samples was used in these experiments. After generating the complementary DNA, PCR was performed with specific primer for mouse eNOS (forward primer, 5′-GAGATCACTGAGCTCTGTATCCAAC-3′; reverse primer, 5′-CTCATTTTCCAGGTGCTTCATGAAG-3′). Conditions for the PCR were as follows: 95 °C for 4 min; 95 °C for 30 s, 61 °C for 30 s, and 72 °C for 30 s (40 cycles); 72 °C for 7 min followed by 4 °C. PCR product was purified and confirmed by sequencing.

***In situ* hybridization.** Frozen sections (10 µm) on 3-aminopropyltriethoxysilane-coated slides were fixed with 4% paraformaldehyde in PBS for 15 min. After being washed three times with PBS, sections were treated with 0.2 M HCl for 15 min at room temperature, followed by incubation with prewarmed proteinase K (20 µg ml^{-1}) for 15 min. The sections were then refixed in 4% paraformaldehyde in PBS. After being washed with PBS, sections were prehybridized with hybridization buffer (50% formamide, 2 × SSC, 50 mM phosphate buffer pH 7.0, 1 × Denhardt's solution, 5% dextran); hybridized with digoxigenin-labelled (Roche) probe in 1 µg ml^{-1} hybridization buffer, overnight at 37 °C. The next day, hybridized sections were washed with wash buffer (100 mM Tris-HCl

pH 7.5, 150 mM NaCl) for 10 min followed by 30 min incubation with 1 × blocking buffer (Roche). After blocking, sections were incubated with 1:5,000 diluted anti-digoxigenin antibody (Roche) for 30 min, washed with wash buffer twice for 15 min and equilibrated with detection buffer (100 mM Tris-HCl pH 9.5, 100 mM NaCl) for 3 min. Subsequently, sections were incubated with colour substrate (Roche) in detection buffer and the reaction was stopped with TE buffer. Sections were observed by light microscopy.

Statistical analysis. Data are expressed as means ± s.e.m. A Mann–Whitney test (two-tailed) for analysis of human samples or an analysis of variance followed by Bonferroni/Dunn's test for multiple comparisons of mouse samples was used to determine significant. Statistical significance was defined as $P < 0.05$. Statview 5.0 was used for statistical analysis.

31. Sugimoto, H. *et al.* Bone-marrow-derived stem cells repair basement membrane collagen defects and reverse genetic kidney disease. *Proc. Natl Acad. Sci. USA* **103,** 7321–7326 (2006).
32. Walzer, T. *et al.* Identification, activation, and selective *in vivo* ablation of mouse NK cells via NKp46. *Proc. Natl Acad. Sci. USA* **104,** 3384–3389 (2007).
33. Tanaka, T. S. *et al.* Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc. Natl Acad. Sci. USA* **97,** 9127–9132 (2000).

LETTERS

Crucial role for the Nalp3 inflammasome in the immunostimulatory properties of aluminium adjuvants

Stephanie C. Eisenbarth[1,2], Oscar R. Colegio[1,3], William O'Connor Jr[1], Fayyaz S. Sutterwala[1,5] & Richard A. Flavell[1,4]

Aluminium adjuvants, typically referred to as 'alum', are the most commonly used adjuvants in human and animal vaccines worldwide, yet the mechanism underlying the stimulation of the immune system by alum remains unknown. Toll-like receptors are critical in sensing infections and are therefore common targets of various adjuvants used in immunological studies. Although alum is known to induce the production of proinflammatory cytokines *in vitro*, it has been repeatedly demonstrated that alum does not require intact Toll-like receptor signalling to activate the immune system[1,2]. Here we show that aluminium adjuvants activate an intracellular innate immune response system called the Nalp3 (also known as cryopyrin, CIAS1 or NLRP3) inflammasome. Production of the pro-inflammatory cytokines interleukin-1β and interleukin-18 by macrophages in response to alum *in vitro* required intact inflammasome signalling. Furthermore, *in vivo*, mice deficient in Nalp3, ASC (apoptosis-associated speck-like protein containing a caspase recruitment domain) or caspase-1 failed to mount a significant antibody response to an antigen administered with aluminium adjuvants, whereas the response to complete Freund's adjuvant remained intact. We identify the Nalp3 inflammasome as a crucial element in the adjuvant effect of aluminium adjuvants; in addition, we show that the innate inflammasome pathway can direct a humoral adaptive immune response. This is likely to affect how we design effective, but safe, adjuvants in the future.

Shortly after the discovery that alum could be used as an adjuvant in the 1920s (ref. 3), a hypothesis was put forth that alum stimulated an immune response by acting as a 'depot'; antigens were proposed to be slowly released in a particulate form that was favourable for uptake by antigen-presenting cells (APCs), thereby enhancing the immune response to the antigen (reviewed in ref. 4). Since then, many of the signals used by APCs to initiate T-cell responses have been identified along with the immune stimuli (for example, Toll-like receptor (TLR) ligands) required to enhance interactions between APCs and T cells. However, the cellular signalling pathways triggered by alum that induce effective immunity against antigens have remained elusive.

The initiation of adaptive immune responses is controlled by innate immune signals. Regulation of these immune signals relies on a large group of intracellular and extracellular receptors called pattern recognition receptors[5]. The best-described class of these receptors is the TLRs, which sense conserved molecular patterns from a wide range of microbes. Whereas TLRs sense non-self motifs of infectious organisms, another class of intracellular pattern recognition receptors, the NOD-like receptors (NLRs), can sense stimuli of microbial origin as well as endogenous markers of cellular damage (for example ATP or uric acid crystals)[6,7]. Nalp3, a member of the NLR family, along with ASC (also known as Pycard) and caspase-1, forms a molecular platform called the inflammasome, which regulates the cleavage and release of the potent pro-inflammatory cytokines interleukin (IL)-1β, IL-18 and IL-33 (ref. 8). One recently described endogenous molecule that activates the Nalp3 inflammasome is crystalline (but not soluble) uric acid (monosodium urate; MSU)[9–11].

Aluminium particles of various aluminium adjuvants form insoluble particles that can aggregate, are readily phagocytosed by macrophages and have been shown to stimulate IL-1β and IL-18 production *in vitro*[12–15]. We formed the hypothesis that the particulate nature of alum might be recognized by NLRs, much like crystalline MSU. To test whether alum activates the Nalp3 inflammasome, we used primary peritoneal macrophages from mice deficient in critical signalling components of the Nalp3 inflammasome. Because inflammasome activation requires two signals for the production of mature IL-1β, we first primed macrophages with lipopolysaccharide (LPS) and then exposed them to aluminium adjuvants. Consistent with previous reports[13–15], aluminium adjuvants induced the production of IL-1β and IL-18 from wild-type (C57BL/6; WT) primary murine macrophages (Fig. 1a, d), bone-marrow-derived macrophages (Supplementary Fig. 1a) and bone-marrow-derived dendritic cells (Supplementary Fig. 1b) *in vitro*. IL-1β secretion was dependent on the dose of alum (Fig. 1b) and peaked between 8 and 10 h of stimulation with alum in WT macrophages, but continued out to 48 h (Fig. 2c and data not shown).

In contrast, macrophages from animals deficient in Nalp3, ASC or caspase-1 failed to produce IL-1β or IL-18 on stimulation with multiple types of aluminium adjuvant (Fig. 1c, d, Supplementary Fig. 1b and data not shown). Another member of the NLR family, Ipaf (also known at NLRC4), also forms an inflammasome, which is capable of activating caspase-1 in response to several different Gram-negative bacteria[16,17]. Ipaf-deficient macrophages were fully capable of secreting IL-1β in response to LPS and alum (Fig. 1c), suggesting that alum-induced IL-1β secretion is specifically dependent on the Nalp3 inflammasome. Consistent with NLR, but not TLR activation, alum did not induce the production of IL-6 or tumour necrosis factor (TNF)-α by primary macrophages *in vitro* (Supplementary Fig. 2a, b). Neither of two other common adjuvants, complete Freund's adjuvant (CFA) and incomplete Freund's adjuvant (IFA), induced IL-1β production by macrophages (Fig. 1e).

Caspase-1 activation involves autocatalytic processing of the 45-kDa pro-caspase-1 to generate two subunits, p20 and p10. Caspase-1 activation in LPS-primed WT macrophages stimulated with alum was detected by western blotting by the appearance of the p10 cleavage product 4 h after the addition of alum (Fig. 2a). Consistent with the lack of IL-1β production, caspase-1 activation was absent in macrophages deficient in Nalp3 and ASC that were exposed to LPS and alum (Fig. 2a, b). Nalp3 knockout macrophages did not show caspase-1 activation or IL-1β production even at later time points,

[1]Department of Immunobiology, [2]Department of Laboratory Medicine, [3]Department of Dermatology, and [4]Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [5]Inflammation Program, Department of Medicine, University of Iowa, Iowa City, Iowa 52241, USA.

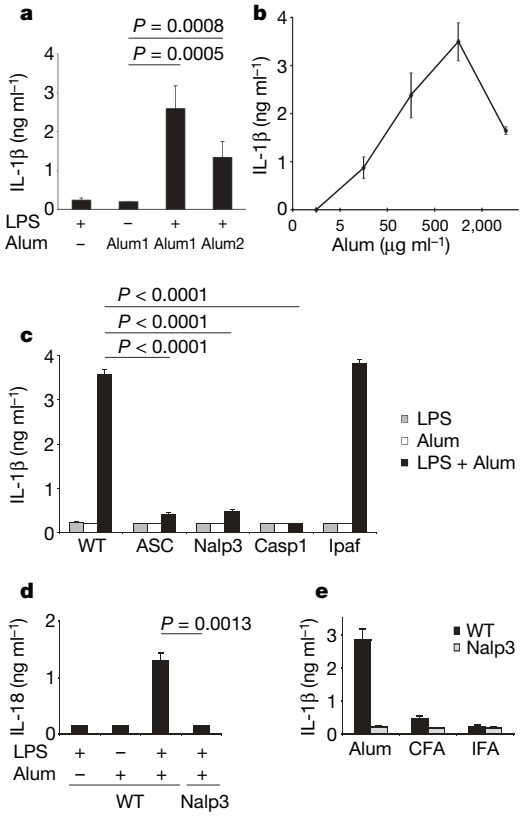


Figure 1 | Aluminium-containing adjuvants stimulate macrophages to produce the pro-inflammatory cytokines IL-1β and IL-18 in a Nalp3 inflammasome-dependent manner. **a**, Macrophages were stimulated with 50 ng ml^{-1} LPS for 18 h and then 500 μg ml^{-1} Imject alum ('Alum1') or aluminium hydroxide gel ('Alum2') for 8 h. IL-1β released into culture supernatants was measured by ELISA with a minimum detection level of 200 pg ml^{-1}. **b**, LPS-primed macrophages were stimulated with the indicated amount of Imject alum for 8 h and analysed as in **a**. **c**, Unprimed or LPS-primed WT, ASC-deficient, Nalp3-deficient, caspase-1-deficient (Casp1) and Ipaf-deficient macrophages were stimulated with Imject alum (500 μg ml^{-1}) for 8 h, and the IL-1β released into the culture supernatants was measured by ELISA. **d**, WT or Nalp3-deficient macrophages were stimulated as in **c**, and the IL-18 released was measured by ELISA. **e**, WT or Nalp3-deficient LPS-primed macrophages were stimulated with either Imject alum, CFA (120 μg ml^{-1}) or IFA (30 μl ml^{-1}) for 8 h and analysed as in **a**. Determinations were performed in triplicate and are expressed as means and s.d.; data are from one of at least three independent experiments.

arguing against delayed caspase-1 activation by alum in the absence of Nalp3 (Fig. 2a, c). These data demonstrate that alum activates macrophages *in vitro* to secrete mature IL-1β in a manner dependent on the Nalp3 inflammasome.

To understand how alum might stimulate the inflammasome pathway, we first tested whether the endocytic ability of macrophages was required for the alum-stimulated production of IL-1β. Inhibiting actin or tubulin polymerization with either cytochalasin B or colchicine, respectively, inhibited IL-1β production by LPS and alum (Fig. 3a) but did not affect secretion of the inflammasome-independent cytokines TNF-α or IL-6 (Supplementary Fig. 2a, b). Neither cytochalasin B nor colchicine decreased IL-1β production in response to stimulation with ATP, which uses the P2X7 receptor (P2X7R) to activate the Nalp3 inflammasome[18,19], confirming that macrophages were still viable and capable of secreting inflammasome-dependent IL-1β (Fig. 3a).

ATP and MSU released from dying and injured cells into the extracellular milieu may activate the Nalp3 inflammasome[8,9,19]. *In vitro*, alum induced cell death at very high doses (that did not induce significant IL-1β) in WT macrophages and in macrophages deficient in Nalp3 and Caspase-1 (Fig. 3b); however, the induction of IL-1β by alum did not depend on the presence of MSU because the addition of uricase, which degrades MSU crystals and prevents the induction of IL-1β (ref. 10), had no effect on IL-1β production in response to LPS and alum (Fig. 3c). To exclude the possibility that Nalp3 inflammasome activation was in response to ATP release caused by alum-induced cellular damage, we used macrophages from P2X7R knockout mice. P2X7R-deficient macrophages showed no defect in IL-1β production after stimulation with LPS and alum (Fig. 3d). Taken together, these data support a model of active endocytosis of alum by viable macrophages leading to Nalp3 inflammasome activation with the resultant secretion of pro-inflammatory cytokines.

Although several diverse stimuli activate the Nalp3 inflammasome, the efflux of cellular potassium seems to be a common step shared by these stimuli and is required for Nalp3-dependent caspase-1 activation; it has therefore been suggested that the inflammasome acts as a sensor of cellular membrane disruption[7]. Preventing this potassium efflux by increasing extracellular potassium inhibits inflammasome activation with a variety of Nalp3 triggers[20]. Indeed, increased extracellular potassium significantly inhibited alum-induced IL-1β production from macrophages (Fig. 3e) but not the LPS-dependent production of TNF-α or IL-6 (data not shown). Pannexin pores are thought to have a function in inflammasome activation induced by ATP, nigericin or maitotoxin, possibly by facilitating potassium efflux[21]. We did not detect a significant difference in IL-1β secretion between macrophages exposed to a pannexin-pore-blocking peptide and those exposed to a

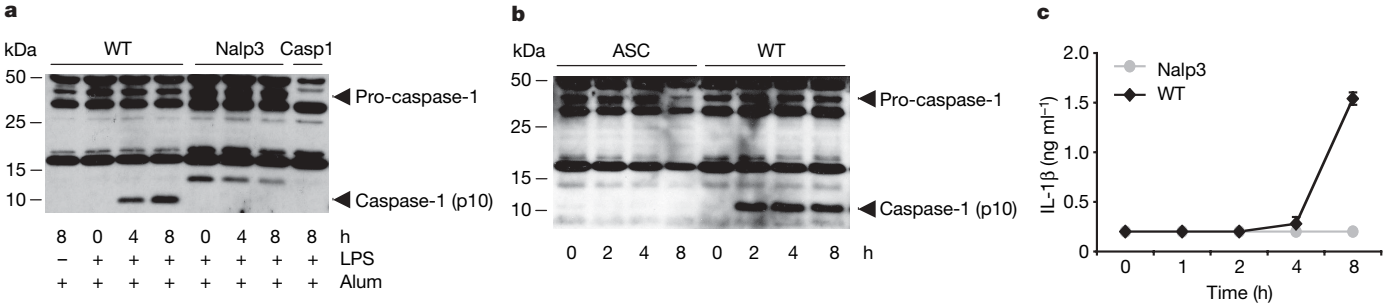


Figure 2 | Caspase-1 activation by aluminium adjuvants requires Nalp3 and ASC. **a**, **b**, LPS-primed or unprimed macrophages from WT mice (**a**, **b**), Nalp3-deficient mice and caspase-1-deficient (Casp1) mice (**a**) and from ASC-deficient mice (**b**) were stimulated with Imject alum (500 μg ml^{-1}) for the indicated durations, and cell lysates were immunoblotted with antibodies against the p10 subunit of caspase-1. Macrophages from caspase-1 knockout mice were stimulated as indicated to provide a reference for some of the non-specific bands seen with this antibody. Indeed, the band at 15 kDa in the Nalp3 knockout samples in **a** and in all samples in **b** is distinct from the p10 band representing active caspase-1. **c**, LPS-primed macrophages from WT or Nalp3-deficient mice were stimulated with Imject alum (500 μg ml^{-1}) for the indicated durations, and the IL-1β released into culture supernatants was measured by ELISA. Results are shown as means ± s.d. from one of three independent experiments.

scrambled peptide. Therefore we do not currently have evidence that pannexin pores mediate alum-induced inflammasome activation. Alum is therefore a new Nalp3 trigger and, like other triggers, may induce inflammasome activation through membrane disruption with resultant potassium efflux.

Aluminium adjuvants are used in human vaccines to induce a potent humoral response; alum is also used as a potent adjuvant to induce T helper type 2 (T_H2)-mediated inflammation in murine allergy/asthma models. Given the Nalp3-dependent activation of macrophages that we observed *in vitro*, we tested whether immunity in mice against a model protein antigen, ovalbumin, required a functional Nalp3 inflammasome. Ovalbumin-specific IgG1 antibody induction was significantly decreased in Nalp3-deficient, ASC-deficient and caspase-1-deficient mice primed intraperitoneally with ovalbumin and alum (Fig. 4a) or subcutaneously with another protein antigen, human serum albumin (HSA) in alum (Supplementary Fig. 3), but was not affected in MyD88 knockout mice (Supplementary Fig. 4; ref. 2). We tested whether Nalp3 and ASC knockout mice have a general antibody-production defect by immunizing them with the adjuvant CFA. Ovalbumin-specific IgG2c (Fig. 4b) and IgG1 (not shown) in Nalp3 and ASC knockout mice were equivalent to levels in WT mice but, as expected, CFA-induced ovalbumin-specific IgG2c was completely dependent on MyD88 (Fig. 4b).

T_H2 cell priming was also impaired in Nalp3, ASC and caspase-1 knockout mice as demonstrated by decreased airway eosinophilia and hilar lymph-node IL-5 production in an alum-dependent model of asthma (Fig. 4c, d). The overall inflammation was decreased in these knockout mice without evidence of a switch to a T_H1 response (typically characterized by airway neutrophilia and IgG2c induction). Consistent with previous reports, alum-induced T_H2 responses are not affected in mice lacking MyD88 (ref. 2) or lacking both MyD88 and TRIF (ref. 1; Fig. 4c and Supplementary Fig. 4). Previous studies have suggested that antigen must be physically associated with (although not necessarily adsorbed on) alum for it to have an adjuvant effect[22]. Indeed, we saw a significantly impaired antibody response (Supplementary Fig. 5a) and an absence of T_H2 inflammation in the airways when alum and ovalbumin were injected separately into the peritoneum (Supplementary Fig. 5b). In mouse cells, but not in human cells, there is a clear requirement *in vitro* for two signals to activate the inflammasome and to produce pro-IL-1β (LPS and alum), yet it is not clear what is providing the first signal for alum *in vivo* (or other Nalp3 stimuli including MSU). We have preliminary evidence from *in vitro* studies that IL-1β itself can prime macrophages for alum-induced inflammasome activation (data not

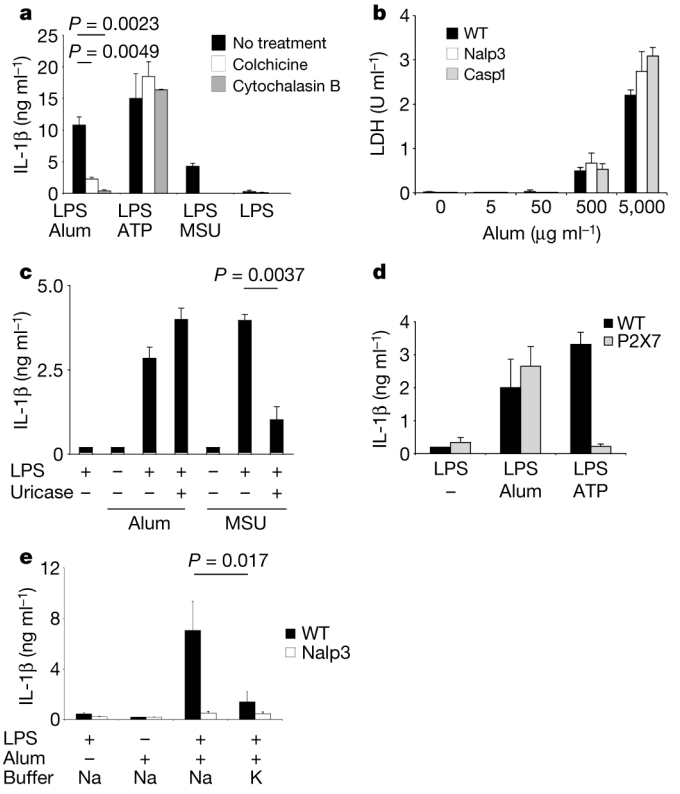


Figure 3 | Alum requires intact endocytic macrophage machinery and causes potassium-gradient-dependent IL-1β secretion without causing significant cell death. **a**, LPS-primed peritoneal macrophages were treated with either colchicine (28 μg ml^{-1}) or cytochalasin B (10 μM) for 1 h before the addition of Imject alum (500 μg ml^{-1}), ATP (5 mM) or MSU (200 μg ml^{-1}). **b**, Lactate dehydrogenase (LDH) release was measured from LPS-primed WT, Nalp3-deficient and caspase-1-deficient (Casp1) macrophage culture supernatants stimulated with the indicated amounts of Imject alum. **c**, Unprimed or LPS-primed WT macrophages were stimulated for 8 h with either Imject alum (500 μg ml^{-1}) or MSU (200 μg ml^{-1}) in the presence or absence of 2 U ml^{-1} uricase. **d**, LPS-primed macrophages from WT or P2X7R-deficient (P2X7) mice were stimulated with Imject alum (500 μg ml^{-1}) or ATP (5 mM) and samples were analysed as in **a**. **e**, Unprimed or LPS-primed WT or Nalp3-deficient macrophages were stimulated with Imject alum in serum-free buffer with either 150 mM NaCl or 150 mM KCl and analysed as in **a**. Determinations were performed in triplicate and are expressed as means and s.d.; data are from one of at least three independent experiments.

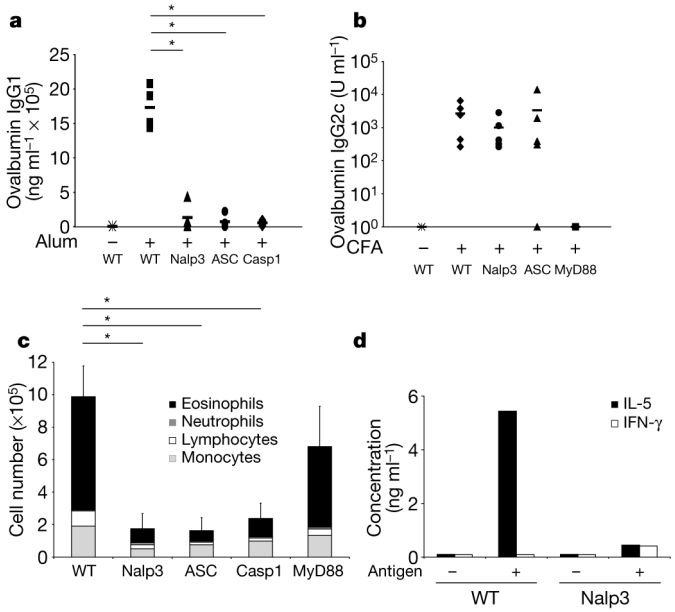


Figure 4 | Antibody production and T_H2-dependent inflammation induced by aluminium adjuvants are decreased in the absence of Nalp3, ASC and caspase-1. **a**, WT, Nalp3-deficient, ASC-deficient or caspase-1-deficient (Casp1) mice (three to five mice per group) 6–8 weeks old were injected intraperitoneally with ovalbumin adsorbed on Imject alum on day 0 and again on day 10. Mice were challenged intranasally with ovalbumin on days 21, 22 and 23. Sera were collected from mice on day 25 and analysed for ovalbumin-specific IgG1 by ELISA as described previously[2]. Asterisk, $P < 0.03$; nonparametric Mann–Whitney *U*-test. **b**, WT, Nalp3-deficient, ASC-deficient or MyD88-deficient mice (three to five mice per group) were primed subcutaneously with ovalbumin in CFA on day 0 and on day 10 in IFA. Sera were collected on day 21 and analysed for ovalbumin-specific IgG2c by ELISA. **c**, Three to five mice per group were primed and challenged as in **a**; bronchoalveolar lavage was collected on day 25 and analysed as described previously[2] (see Methods) (total cell number; means and s.d. are shown). Asterisk, $P < 0.03$; nonparametric Mann–Whitney *U*-test. **d**, Lung draining (hilar) lymph nodes were collected from WT and Nalp3-deficient mice primed and challenged as in **a** and pooled within each group for restimulation; cells were restimulated *in vitro* with (+) or without (−) 200 μg ml^{-1} ovalbumin and mitomycin-C-treated splenocytes for 48 h. Supernatants were analysed for IL-5 (filled bars) or IFN-γ (open bars).

shown); these results are consistent with previous reports that IL-1β can act in an autocrine manner to induce its own gene expression[23]. Other groups have similarly seen macrophage priming with cytokines (for example TNF-α) instead of LPS[13]. Combining the above information with the fact that alum must be encountered simultaneously with antigen *in vivo* for efficient priming suggests that the antigen might provide the first signal either directly, or indirectly by inciting the production of local pro-inflammatory cytokines from resident monocytes or specialized cells recruited by alum[24]. Once the first signal has primed the cell, alum provides the second signal for activation of the Nalp3 inflammasome. These two stimuli must be sensed by the same cell for effective immune activation, thereby increasing the specificity of an immune response and perhaps explaining why alum (which readily adsorbs antigens) is such an effective adjuvant.

Thus, by eliminating signalling through the Nalp3 inflammasome, we have eliminated one critical pathway used by alum to initiate humoral and cellular immunity. In doing so, aluminium hydroxide adjuvants 'hijack' an innate immune pathway that is exquisitely sensitive to cellular damage, perhaps as a result of the similarity to MSU in its physical structure. Although intraperitoneal MSU induces peritonitis[9] and subcutaneous MSU in concert with antigen injection has been used *in vivo* to initiate CD8 T-cell responses[10], we predicted, on the basis of our findings, that this Nalp3 stimulant would also induce a significant antibody response to a protein antigen. Indeed, MSU injected intraperitoneally with antigen induces an IgG1-type antibody response similar in nature to that induced by alum in WT mice but not in Nalp3-deficient mice (Supplementary Fig. 6). These mice did not develop a significant T_H1-type antibody response (IgG2c) under these immunization conditions (data not shown), suggesting that MSU and alum induce a similar pattern of inflammation when injected at similar doses in the same location.

A critical question regarding the mechanism by which alum influences immunity is how alum initiates lymphocyte activation and how it favours T_H2 differentiation over T_H1 differentiation. Inflammasome-dependent cytokines have been implicated in various aspects of T_H2 responses: IL-1 has classically been thought to promote T_H2 cell proliferation[25,26], IL-33 is a potent pro-T_H2 stimulus[26] and IL-18 has been shown to augment IgE antibody production (although it primarily potentiates T_H1 responses)[27]. On the basis of our *in vitro* findings, we would predict that local production of IL-1β, IL-18 and/or IL-33 could induce the requisite signals for activation of the adaptive immune system. Indeed, we found a lower expression of *Il1b* mRNA from peritoneal cells of Nalp3-deficient mice immunized with ovalbumin and alum than in WT mice (Supplementary Fig. 7). In further support of an IL-1-dependent model, the antibody response in another immunization model has been shown to be defective in IL-1α/IL-1β knockout mice as the result of a defect in the induction of CD40L on T cells by activated APCs[28]. However, there is no antibody production or T_H2 defect after alum priming in MyD88 knockout mice. MyD88 is critical in the IL-1 receptor signalling cascade[29], although one recent study has identified a MyD88-independent IL-1 pathway[30]. It will therefore be of interest to study the relative roles of IL-1 family members in alum-dependent priming in future work. In addition, as new functions of caspase-1 and the inflammasome are uncovered, we will further understand how stimulation of this potent pro-inflammatory machinery results in activation of the adaptive immune response.

METHODS SUMMARY

Mice. The generation of mice deficient in Nalp3, ASC, Ipaf, caspase-1 and P2X7R has been reported previously[8,16,18]. Nalp3-deficient, Caspase-1-deficient and ASC-deficient mice were backcrossed nine generations, and Ipaf-deficient mice were backcrossed six generations onto a C57BL/6 background. Age-matched and sex-matched C57BL/6 mice from the National Cancer Institute were used as all WT controls. All protocols used in this study were approved by the Yale Institutional Animal Care and Use Committee.

Macrophages. The generation of thioglycollate-elicited peritoneal and bone-marrow-derived macrophages and bone-marrow-derived dendritic cells has been described previously[2,8]. Unless indicated, macrophages were primed by stimulating with 50 ng ml^{-1} LPS from *Escherichia coli* serotype 0111:B4 (Invivogen) for 16–18 h before stimulation with Imject alum (unless otherwise indicated), MSU or ATP. For ATP-stimulated cells, the medium was changed at 20 min and all stimulants were replaced. Macrophage cell death was measured by the release of lactate dehydrogenase with a cytotoxicity detection kit (Promega).

Sensitizations. For intraperitoneal sensitization, 6–8-week-old mice were injected intraperitoneally on day 0 with 50 μg of ovalbumin (Grade V; Sigma) adsorbed on 4 mg of Imject alum and again on day 10 with 25 μg of ovalbumin adsorbed on 4 mg of Imject alum. Mice were challenged intranasally with 25 μg of ovalbumin in PBS on days 21, 22 and 23. Mice were killed for analysis on day 25. For subcutaneous sensitization, mice were injected subcutaneously on day 0 with 50 μg of ovalbumin in 400 μg (180 μl) of CFA and again on day 10 with 25 μg of ovalbumin in 180 μl of IFA.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 March; accepted 1 April 2008.
Published online 21 May 2008.

1. Gavin, A. L. *et al.* Adjuvant-enhanced antibody responses in the absence of toll-like receptor signaling. *Science* **314,** 1936–1938 (2006).
2. Piggott, D. A. *et al.* MyD88-dependent induction of allergic Th2 responses to intranasal antigen. *J. Clin. Invest.* **115,** 459–467 (2005).
3. Glenny, A. T., Pope, C. G., Waddington, H. & Wallace, U. The antigenic value of toxoid precipitated by potassium alum. *J. Pathol. Bacteriol.* **29,** 31–40 (1926).
4. Lindblad, E. B. Aluminium compounds for use in vaccines. *Immunol. Cell Biol.* **82,** 497–505 (2004).
5. Medzhitov, R. Recognition of microorganisms and activation of the immune response. *Nature* **449,** 819–826 (2007).
6. Mariathasan, S. & Monack, D. M. Inflammasome adaptors and sensors: intracellular regulators of infection and inflammation. *Nature Rev. Immunol.* **7,** 31–40 (2007).
7. Sutterwala, F. S., Ogura, Y. & Flavell, R. A. The inflammasome in pathogen recognition and inflammation. *J. Leukoc. Biol.* **82,** 259–264 (2007).
8. Sutterwala, F. S. *et al.* Critical role for NALP3/CIAS1/Cryopyrin in innate and adaptive immunity through its regulation of caspase-1. *Immunity* **24,** 317–327 (2006).
9. Martinon, F., Petrilli, V., Mayor, A., Tardivel, A. & Tschopp, J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440,** 237–241 (2006).
10. Shi, Y., Evans, J. E. & Rock, K. L. Molecular identification of a danger signal that alerts the immune system to dying cells. *Nature* **425,** 516–521 (2003).
11. Chen, C. J. *et al.* Identification of a key pathway required for the sterile inflammatory response triggered by dying cells. *Nature Med.* **13,** 851–856 (2007).
12. Rimaniol, A. C. *et al.* Aluminum hydroxide adjuvant induces macrophage differentiation towards a specialized antigen-presenting cell type. *Vaccine* **22,** 3127–3135 (2004).
13. Li, H., Nookala, S. & Re, F. Aluminum hydroxide adjuvants activate caspase-1 and induce IL-1β and IL-18 release. *J. Immunol.* **178,** 5271–5276 (2007).
14. Mannhalter, J. W., Neychev, H. O., Zlabinger, G. J., Ahmad, R. & Eibl, M. M. Modulation of the human immune response by the non-toxic and non-pyrogenic adjuvant aluminium hydroxide: effect on antigen uptake and antigen presentation. *Clin. Exp. Immunol.* **61,** 143–151 (1985).
15. Sokolovska, A., Hem, S. L. & HogenEsch, H. Activation of dendritic cells and induction of $CD4^+$ T cell differentiation by aluminum-containing adjuvants. *Vaccine* **25,** 4575–4585 (2007).
16. Sutterwala, F. S. *et al.* Immune recognition of *Pseudomonas aeruginosa* mediated by the IPAF/NLRC4 inflammasome. *J. Exp. Med.* **204,** 3235–3245 (2007).
17. Mariathasan, S. *et al.* Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. *Nature* **430,** 213–218 (2004).
18. Solle, M. *et al.* Altered cytokine production in mice lacking $P2X_7$ receptors. *J. Biol. Chem.* **276,** 125–132 (2001).
19. Mariathasan, S. *et al.* Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440,** 228–232 (2006).
20. Petrilli, V. *et al.* Activation of the NALP3 inflammasome is triggered by low intracellular potassium concentration. *Cell Death Differ.* **14,** 1583–1589 (2007).
21. Pelegrin, P. & Surprenant, A. Pannexin-1 couples to maitotoxin- and nigericin-induced interleukin-1β release through a dye uptake-independent pathway. *J. Biol. Chem.* **282,** 2386–2394 (2007).
22. Chang, M. *et al.* Degree of antigen adsorption in the vaccine or interstitial fluid and its effect on the antibody response in rabbits. *Vaccine* **19,** 2884–2889 (2001).
23. Toda, Y. *et al.* Autocrine induction of the human pro-IL-1β gene promoter by IL-1β in monocytes. *J. Immunol.* **168,** 1984–1991 (2002).
24. Jordan, M. B., Mills, D. M., Kappler, J., Marrack, P. & Cambier, J. C. Promotion of B cell immune responses via an alum-induced myeloid cell population. *Science* **304,** 1808–1810 (2004).
25. Kaye, J. *et al.* Growth of a cloned helper T cell line induced by a monoclonal antibody specific for the antigen receptor: interleukin 1 is required for the expression of receptors for interleukin 2. *J. Immunol.* **133,** 1339–1345 (1984).

26. Dunne, A. & O'Neill, L. A. The interleukin-1 receptor/Toll-like receptor superfamily: signal transduction during inflammation and host defense. *Sci. STKE* **2003,** re3 (2003).
27. Yoshimoto, T. *et al.* IL-18 induction of IgE: dependence on $CD4^+$ T cells, IL-4 and STAT6. *Nature Immunol.* **1,** 132–137 (2000).
28. Nakae, S., Asano, M., Horai, R., Sakaguchi, N. & Iwakura, Y. IL-1 enhances T cell-dependent antibody production through induction of CD40 ligand and OX40 on T cells. *J. Immunol.* **167,** 90–97 (2001).
29. Adachi, O. *et al.* Targeted disruption of the MyD88 gene results in loss of IL-1- and IL-18-mediated function. *Immunity* **9,** 143–150 (1998).
30. Davis, C. N. *et al.* MyD88-dependent and -independent signaling by IL-1 in neurons probed by bifunctional Toll/IL-1 receptor domain/BB-loop mimetics. *Proc. Natl Acad. Sci. USA* **103,** 2953–2958 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Zenewicz, Y. Ogura, A. Williams and Y. Wan for discussion and review of this manuscript; A. Coyle, E. Grant and J. Bertin for providing ASC-deficient, Nalp3-deficient and Ipaf-deficient mice; and J. Genzen for providing the P2X7R-deficient mice. This work was supported by the Ellison Foundation, the Bill and Melinda Gates Foundation through the Grand Challenges in Global Health Initiative, and National Institutes of Health grant K08 (F.S.S.). R.A.F. is an Investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.A.F. (richard.flavell@yale.edu).

METHODS

Materials. All reagents were purchased from Sigma except colchicine (Calbiochem), monosodium urate crystals (Alexis), Imject alum (Pierce), CFA (Difco) and IFA (Difco). Antibody pairs for ELISA were purchased from R&D Systems (IL-1β), MBL (IL-18), BD Pharmingen (IL-5, IFN-γ and IL-6) or from eBioscience (TNF-α). Ovalbumin-specific IgG1 and IgG2c were measured by ELISA as described previously[2], and secondary antibodies were purchased from BD Pharmingen. HSA-specific IgG1 was performed as above, with HSA used for coating instead of ovalbumin.

Western blotting. Electrophoresis of proteins was performed with the NuPAGE system (Invitrogen) in accordance with the manufacturer's protocol. In brief, at the indicated time after alum addition, LPS-primed macrophages were lysed in lysis buffer (50 mM Tris-HCl pH 8.0, 5 mM EDTA, 150 mM NaCl, 1% Triton-X100 and a protease inhibitor cocktail (Roche)) and stored at −80 °C until analysed. Proteins were separated on a NuPAGE gel and transferred to a PVDF (poly(vinylidene difluoride)) membrane by electroblotting. To detect caspase-1, a rabbit polyclonal anti-mouse caspase-1 p10 antibody (clone M-20; Santa Cruz Biotechnology) was used.

Statistical analysis. We performed statistical analysis by using an unpaired Student's *t*-test for all studies unless otherwise indicated. We considered $P < 0.05$ to be statistically significant. For Fig. 4a, c and Supplementary Figs 5 and 6 we performed a nonparametric Mann–Whitney *U*-test.

Inhibition of potassium efflux. To inhibit potassium efflux, macrophages were primed with LPS for 18 h and then the medium was replaced with a serum-free buffer containing 150 mM KCl with the following composition: 10 mM HEPES, 5 mM NaH_2PO_4, 150 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1% BSA, pH adjusted to 7.4 with KOH. For comparison, a buffer with 150 mM sodium chloride was used: 10 mM HEPES, 150 mM NaCl, 5 mM KH_2PO_4, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 1% BSA, pH adjusted to 7.4 with NaOH.

Bronchoalveolar lavage analysis. Mice were primed and challenged as indicated. On the day of analysis, mice were killed and bronchoalveolar lavage was performed as described previously[2]. In brief, inflammatory cells in bronchoalveolar lavage fluid were obtained by cannulation of the trachea and lavage of the airway lumen with PBS. Red blood cells were lysed in bronchoalveolar lavage fluid samples, total cell numbers were counted with a haemocytometer and cytospin slides were prepared by haematoxylin and eosin staining with Diff-Quick (Dade Behring Inc.).

Relative gene expression analysis. RNA from cells was isolated with the RNEasy/Qiashredder purification system (Qiagen) in accordance with the manufacturer's protocol. RNA was subjected to reverse transcriptase with Superscript II (Invitrogen) with oligo(dT) primer in accordance with the manufacturer's protocol. cDNA was semi-quantified using commercially available primer/probe sets (Applied Biosystems) and analysed with the $\Delta\Delta C_t$ (change in cycle threshold) method. All results were normalized to *Hprt* quantified in parallel amplification reactions during each PCR quantification. Results are presented as the relative fold induction levels, where control samples are set to an expression index of 1.

LETTERS

Haem homeostasis is regulated by the conserved and concerted functions of HRG-1 proteins

Abbhirami Rajagopal[1], Anita U. Rao[1], Julio Amigo[2], Meng Tian[3], Sanjeev K. Upadhyay[4], Caitlin Hall[1], Suji Uhm[1], M. K. Mathew[4], Mark D. Fleming[3], Barry H. Paw[2], Michael Krause[5] & Iqbal Hamza[1]

Haems are metalloporphyrins that serve as prosthetic groups for various biological processes including respiration, gas sensing, xenobiotic detoxification, cell differentiation, circadian clock control, metabolic reprogramming and microRNA processing[1–4]. With a few exceptions, haem is synthesized by a multistep biosynthetic pathway comprising defined intermediates that are highly conserved throughout evolution[5]. Despite our extensive knowledge of haem biosynthesis and degradation, the cellular pathways and molecules that mediate intracellular haem trafficking are unknown. The experimental setback in identifying haem trafficking pathways has been the inability to dissociate the highly regulated cellular synthesis and degradation of haem from intracellular trafficking events[6]. *Caenorhabditis elegans* and related helminths are natural haem auxotrophs that acquire environmental haem for incorporation into haemoproteins, which have vertebrate orthologues[7]. Here we show, by exploiting this auxotrophy to identify HRG-1 proteins in *C. elegans*, that these proteins are essential for haem homeostasis and normal development in worms and vertebrates. Depletion of *hrg-1*, or its paralogue *hrg-4*, in worms results in the disruption of organismal haem sensing and an abnormal response to haem analogues. HRG-1 and HRG-4 are previously unknown transmembrane proteins, which reside in distinct intracellular compartments. Transient knockdown of *hrg-1* in zebrafish leads to hydrocephalus, yolk tube malformations and, most strikingly, profound defects in erythropoiesis—phenotypes that are fully rescued by worm HRG-1. Human and worm proteins localize together, and bind and transport haem, thus establishing an evolutionarily conserved function for HRG-1. These findings reveal conserved pathways for cellular haem trafficking in animals that define the model for eukaryotic haem transport. Thus, uncovering the mechanisms of haem transport in *C. elegans* may provide insights into human disorders of haem metabolism and reveal new drug targets for developing anthelminthics to combat worm infestations.

In animals, the terminal enzyme in haem synthesis, ferrochelatase, is located on the matrix side of the inner mitochondrial membrane[8]. Most newly synthesized haem must be transported through mitochondrial membranes to haemoproteins found in distinct intracellular membrane compartments[6]. Haem synthesis is regulated at multiple steps by effectors including iron, haem and oxygen to prevent the uncoordinated accumulation of haem or its precursors[5]. *C. elegans* is a haem auxotroph and is therefore a unique genetic animal model in which to identify the molecules and delineate the cellular pathways for eukaryotic haem transport[6]. Haem analogue studies have suggested that a haem uptake system exists in *C. elegans*[7]. Synchronized *C. elegans* cultures grown in axenic mCeHR-2 liquid medium[9] and supplemented with haemin chloride revealed a robust uptake of fluorescent zinc mesoporphyrin IX (ZnMP) at a haem concentration of 20 μM or less, in contrast with worms grown at 100 μM haem or more (Fig. 1a, b), suggesting that the transport and accumulation of haem are regulated.

We conducted genome-wide microarrays to identify genes that are transcriptionally regulated by haem. Wild-type N2 worms were grown for two synchronized generations in 4 μM (low), 20 μM (optimal) and 500 μM (high) haem concentrations in liquid medium and their messenger RNA was hybridized to Affymetrix *C. elegans* genome arrays. Statistical analyses identified changes in 370 genes, of which about 164 had some sequence identity to genes in the human genome databases at the amino-acid level, and more than 90% of the genes had no functional annotation in the *C. elegans* database (Supplementary Table 1).

We postulated that the expression of genes that encode for haem transporters might be elevated during haem deficiency to maximize uptake of dietary haem. To identify candidate haem transporter genes, we sorted the 117 genes to identify those that were specifically upregulated in low haem (Supplementary Table 1, categories 1 and 2) and encoded for proteins with predicted transmembrane domains, transport functions, and/or haem/metal-binding motifs. F36H1.5 was >10-fold upregulated at low haem but was undetectable at 500 μM haem, and the predicted open reading frame of 169 amino acids (≈19 kDa) showed similarities to high-affinity permease transporters[10]. We refer to F36H1.5 as haem responsive gene-4 (*hrg-4*). RNA blotting and qRT–PCR analysis revealed that *hrg-4* mRNA was significantly upregulated (>40-fold) at 4 μM haem but undetectable at 20 and 500 μM haem (Fig. 1c, d). We identified three putative paralogues of *hrg-4* in the *C. elegans* genome; we termed them *hrg-1* (R02E12.6), *hrg-5* (F36H1.9) and *hrg-6* (F36H1.10), with 27%, 39% and 35% overall amino-acid sequence identity, respectively (Fig. 1e and Supplementary Fig. 1a). Although both *hrg-1* and *hrg-4* were highly responsive to haem deficiency (Fig. 1c), the magnitude of change in mRNA expression at 1.5 μM haem and their responsiveness to haem-mediated repression were markedly different (Fig. 1d and inset). By contrast, *hrg-5* and *hrg-6* expression seemed to be constitutive and not regulated by haem (not shown). *hrg-4*, *hrg-5* and *hrg-6* are nematode-specific genes, whereas *hrg-1* has orthologues with about 25% amino acid identity in vertebrates (Fig. 1e, f, and Supplementary Fig. 1). Topology modelling and motif analysis of HRG-1 identified four predicted transmembrane domains (TMDs) and a conserved tyrosine and acidic-dileucine-based sorting signal in the cytoplasmic carboxy terminus (Fig. 1e and Supplementary Fig. 1a)[11]. In addition, residues that could potentially either directly bind haem (H90 in TMD2) or interact with the

[1]Department of Animal & Avian Sciences and Department of Cell Biology & Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA. [2]Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Department of Pathology, Children's Hospital Boston, Boston, Massachusetts 02115, USA. [4]National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK campus, Bangalore 560 065, India. [5]Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.

haem side chains (FARKY) were situated in the C-terminal tail (Fig. 1e, g)[12–14].

To study the function of *hrg-1* genes in haem homeostasis, we generated a *hrg-1::gfp* transcriptional fusion in *C. elegans*. *hrg-1::gfp* was expressed specifically in the intestinal cells in larvae and adults (Fig. 2a). Its expression was regulated by feeding transgenic worms sequentially with *Escherichia coli* that had been grown on agar plates with or without exogenous haem (Fig. 2a). *hrg-1* repression was specific to haem because neither protoporphyrin IX nor iron altered the expression of *hrg-1::gfp* (Fig. 2b). We next assessed the effect of HRG-1 and HRG-4 depletion in worms by RNA-mediated interference with three independent assays: first, the expression of green fluorescent protein (GFP) in the *hrg-1::gfp* haem sensor strain to monitor haem homeostasis; second, the accumulation of fluorescent ZnMP as a function of haem uptake; and third, animal viability in the presence of a cytotoxic haem analogue, gallium protoporphyrin IX (GaPP)[7]. Knockdown of *hrg-4* by RNAi resulted in the expression of *hrg-1::gfp*, even though haem levels sufficient to suppress GFP were present in the diet (Fig. 2c). *hrg-4* RNAi resulted in no detectable accumulation of ZnMP fluorescence in worms that were grown in 1.5 μM haem, a concentration that is sufficient to induce a robust uptake of haem (Fig. 2d). Consistent with these findings was our observation that progeny from *hrg-4* RNAi worms were also markedly resistant to GaPP toxicity (Fig. 2e), in concordance with recent genome-wide studies revealing *hrg-4* expression in the worm intestine[15]. In contrast, *hrg-1* RNAi showed a significant derepression of GFP only at low haem levels in the *hrg-1::gfp* haem sensor strain (Fig. 2c), but no discernible effect on animal viability assessed with GaPP toxicity assays (Fig. 2e). We found that the intensity of ZnMP fluorescence was significantly greater in the intestines of HRG-1-depleted worms than in controls (Fig. 2d and Supplementary Fig. 2). The observed differences in RNAi phenotypes of *hrg-4* and *hrg-1* suggest that haem uptake into worm intestinal cells involves HRG-4, whereas HRG-1 mediates haem homeostasis by means of an intracellular compartment.

To dissect HRG-1 function in a vertebrate genetic model, we used zebrafish (*Danio rerio*). We reasoned that any perturbation in haem homeostasis would be manifested as haematological defects in the fish embryo[16]. BLAST searches revealed an orthologous gene on zebrafish chromosome 6 that shared about 21% amino acid identity with *C. elegans* HRG-1. Whole-mount *in situ* hybridization of zebrafish embryos at the 15-somite stage and 24 h after fertilization showed zebrafish *hrg-1* mRNA expressed throughout the embryo, including the central nervous system (Fig. 3a). To knock down *hrg-1* in zebrafish, antisense morpholinos (MO2) were designed at the splice junctions to selectively induce *hrg-1* mRNA mis-splicing and degradation. Embryos injected with MO2 had severe anaemia and lacked any detectable *o*-dianisidine-positive erythroid cells (Fig. 3b, c). MO2 morphants showed other developmental defects, including hydrocephalus and a curved body with shortened yolk tube.

The phenotypes observed in *hrg-1* knockdown embryos suggested an essential role for zebrafish *hrg-1* in the specification, maintenance or maturation of the erythroid cell lineage. *hrg-1* morphants revealed wild-type levels of *βe1*−globin mRNA, a marker for haemoglobinization in the developing blood island and intermediate cell mass of embryos at 24 h after fertilization, but by 48 h after fertilization there were no detectable globin-producing cells (Supplementary Fig. 3a).

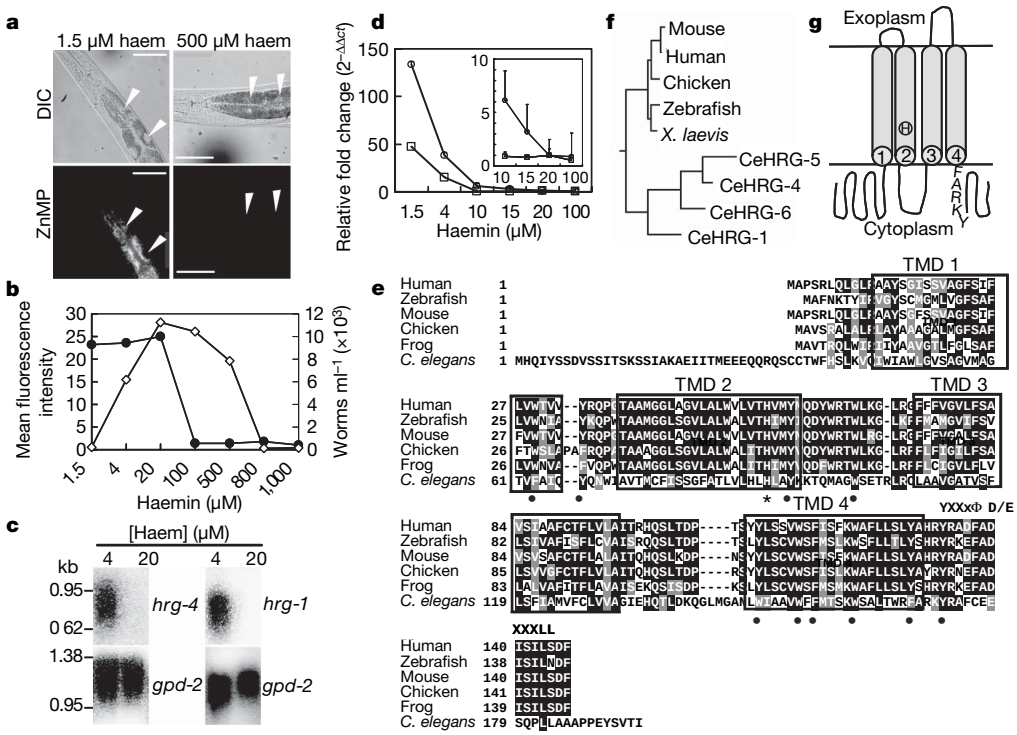


Figure 1 | Identification of *hrg-1* and *hrg-4* in *C. elegans*. **a**, Fluorescent ZnMP (40 μM for 3 h) accumulation in worms grown in mCeHR-2 medium supplemented with 1.5 μM (left) and 500 μM (right) haem. Differential interference contrast (DIC, top) and rhodamine fluorescence (bottom). **b**, Total mean fluorescence intensity (filled circles) of ZnMP accumulated in worms (40 μM for 3 h) after 9 days of growth in mCeHR-2 medium supplemented with the indicated haem concentrations. Open diamonds, growth of worms in haemin. Results are means ± s.d. ($n = 100$). **c**, Northern blot analysis of *hrg-1* and *hrg-4* expression in response to 4 and 20 μM haem in mCeHR-2 medium. The blot was stripped and reprobed with glyceraldehyde 3-phosphate dehydrogenase (*gpd-2*) as loading control. kb, kilobases. **d**, Expression of *hrg-4* (circles) and *hrg-1* (squares) mRNA estimated by quantitative RT–PCR from total RNA obtained from worms grown at the indicated haem concentrations. Each data point shows mean ± s.d. and the results are representative of three separate experiments. Inset: mRNA levels at higher haem concentrations. **e**, Multiple sequence alignment of *C. elegans* HRG-1 with its vertebrate orthologues. Asterisk, histidine (H90); circles, aromatic amino acids; box, putative transmembrane domains; YXXxϕ, C-terminal tyrosine motif; D/EXXxLL, di-leucine motif. **f**, Phylogenetic analysis of HRG-1 proteins using the neighbour-joining method. **g**, Predicted topology of *C. elegans* HRG-1 showing H90 in TMD2, and FARKY, the putative haem-interacting motif, in the cytoplasmic tail.

Moreover, markers for myeloid (*MPO* and *L-plastin*) and thrombocyte (platelet-equivalent, *cd41*) lineages were normal in the *hrg-1* morphant embryos (Supplementary Fig. 3b, c)[17,18]. These findings indicate that zebrafish *hrg-1* is not required for cell lineage specification but rather for maintenance and haemoglobinization of the embryonic erythroid cells. Similarly, *pax 2.1* mRNA expression, a marker of the midbrain/hindbrain boundary organizer, was severely deficient in the central nervous system of MO2 morphants, indicating that midbrain–hindbrain development in zebrafish is also dependent on *hrg-1* (Supplementary Fig. 3d). To verify whether the knockdown phenotypes observed in zebrafish corresponded functionally to the RNAi phenotypes in *C. elegans* (compare Fig. 2c–e with Fig. 3b, c), we co-injected MO2 in the presence and absence of *C. elegans hrg-1* synthetic antisense RNA (cRNA). Despite the modest (21%) sequence identity between the *C. elegans* and zebrafish HRG-1, more than 85% of the morphant embryos were fully rescued by *Cehrg-1* (95 of 108 mutants rescued), in contrast with none for the control embryos (0 of 194 mutants rescued), correcting the defects in anaemia, hydrocephalus and body axis curvature (Fig. 3d, e). These studies suggest that *C. elegans* and zebrafish HRG-1 have a highly conserved function in modulating haem homeostasis.

To dissect the function of HRG-1 in vertebrates further, we examined its gene expression, intracellular localization and biochemical properties in mammalian cells. Genome database searches with *C. elegans* HRG-1 identified an orthologous gene, which we refer to as *hHRG-1* (Fig. 1e), with about 23% and 65% identity to worm and zebrafish HRG-1 proteins, respectively. *hHRG-1* is located on human chromosome 12q13, about 3.2 megabases from *DMT1*, a gene encoding the main iron transporter in mammals[19,20]. RNA blotting of human adult tissues and tumour cell lines detected two *hHRG-1* transcripts about 1.7 and 3.1 kilobases long, with the shorter form predominant (Fig. 4a, b). *hHRG-1* was highly expressed in the brain, kidney, heart and skeletal muscle (Fig. 4a and Supplementary Fig. 4a), and moderately expressed in the liver, lung, placenta and small intestine. *hHRG-1* was abundantly expressed in cell lines derived from duodenum (HuTu 80), kidney (ACHN, HEK-293), bone marrow (HEL, K562) and brain (M17, SH-SY5Y) (Fig. 4b and Supplementary Fig. 4b). Neither altering cellular haem and iron status nor chemically inducing Friend mouse erythroleukaemia (MEL) cells to produce haemoglobin altered *HRG-1* at the transcriptional level in mammalian cells (Supplementary Fig. 5). However, our findings do not exclude the possibility that HRG-1 may be regulated at the post-translational level.

To assess the localization and function of HRG-1 protein, *C. elegans* (Ce)HRG-1, hHRG-1 and CeHRG-4 were tagged at the C terminus with either the haemagglutinin (HA) epitope or GFP variants and transiently transfected into HEK-293 cell lines. All three proteins migrated as a monomer as well as more slowly migrating oligomers (Fig. 4c, lanes 1–6). The oligomerization did not occur in solution after cell lysis or because of protein overexpression, because *in vitro* transcription and translation revealed a single main radiolabelled band corresponding to the monomer for each protein (Fig. 4c, lanes 7–9). Confocal microscopy studies with cells expressing fluorescently tagged proteins showed HRG-4 clearly on the periphery of cells and localized together with a plasma membrane marker. In contrast, CeHRG-1 and hHRG-1 were distributed in an intracellular compartment punctuated throughout the cytoplasm, with about 10% of the total fluorescence on the cell periphery (Fig. 4d). Co-expression of CeHRG-1 and hHRG-1 in the same cell resulted in more than 80% of the two proteins being localized to the same intracellular sites (Fig. 4d). Confocal studies with cellular organelle markers localized CeHRG-1 and hHRG-1 together with LAMP1 (more than 90%), and partly with rab 7 and rab 11 (about 50–70%; Supplementary Fig. 6). These results suggest that HRG-1 proteins are located primarily in endosomes and lysosome-related organelles and are consistent with the presence of sorting motifs in HRG-1 (Fig. 1e)[11].

The presence of conserved haem-binding amino-acid residues in conjunction with the haem-dependent RNAi phenotype in *C. elegans* implied that HRG-1 and HRG-4 may interact with haem. Haemin-agarose affinity chromatography performed on cell lysates from transiently transfected HEK-293 cells showed significant binding of CeHRG-4, CeHRG-1 and hHRG-1 to haem (Fig. 4e), whereas little binding was observed for human ZIP4, an eight-transmembrane-domain zinc transporter that localizes to the plasma membrane and perinuclear cytoplasmic vesicles[21]. Because HRG-1 localized to endosomal–lysosomal organelles whereas HRG-4 is on the plasma

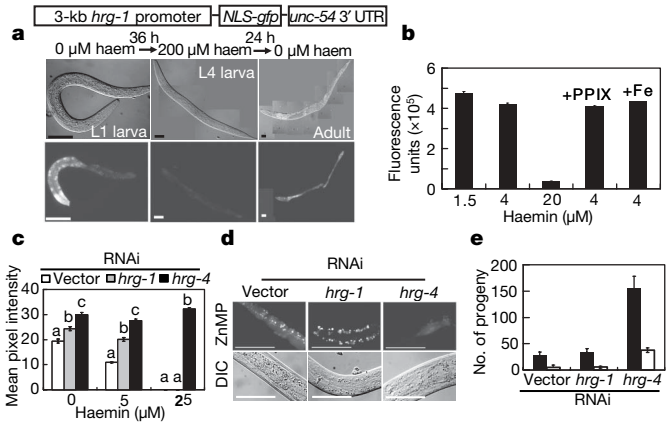


Figure 2 | *hrg-1* and *hrg-4* are essential for haem homeostasis in *C. elegans*. **a**, IQ6011 *hrg-1::gfp* 'haem sensor' strain responds to exogenous haem after sequential exposure to *E. coli* grown on agar plates in the absence (left and right) and presence (middle) of 200 μM haem. UTR, untranslated region. **b**, Spectrofluorometric measurements of GFP in worm lysates from IQ6011 strain grown in the presence of indicated concentrations of haem plus 20 μM protoporphyrin IX (PPIX) or 1 mM $FeCl_3$. Each data point shows mean ± s.d. and the results are representative of three separate experiments. **c**–**e**, Depletion of *hrg-1* or *hrg-4* in worms by RNAi with feeding bacteria. **c**, Dysregulation of GFP (means ± s.e.m.; $n = 35$–45 worms per treatment) in IQ6011 when fed with bacteria grown in the presence of 0, 5 and 25 μM haem. Values with different letter labels are significantly different ($P < 0.001$) within each treatment. **d**, Aberrant ZnMP fluorescence accumulation in worms fed with 10 μM ZnMP for 16 h. Scale bar, 50 μm. **e**, Differences in viable progeny (mean ± s.d.; $n = 30$ P_0 worms per treatment) after 5 days of exposure to 1 μM GaPP plus RNAi bacteria. Filled bars, viable eggs; open bars, larvae. The results for **c**–**e** are representative of at least four separate experiments.

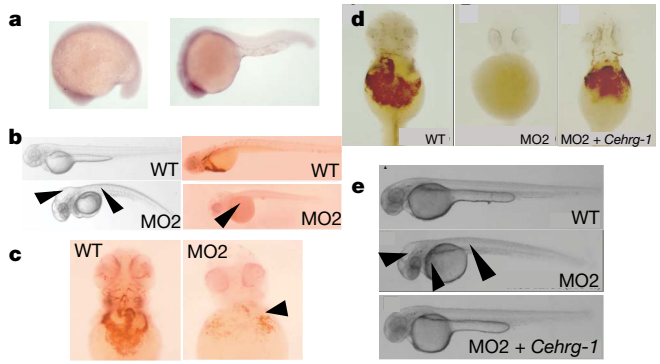


Figure 3 | HRG-1 is essential for erythropoiesis and development in zebrafish. **a**, Zebrafish *hrg-1* expression by whole-mount *in situ* hybridization: left, 15 somites; right, 24 h after fertilization. **b**, Knockdown of zebrafish *hrg-1* by using morpholinos (MO2) against zebrafish *hrg-1* reveals severe anaemia with very few *o*-dianisidine-positive red cells (arrows, right panel), hydrocephalus, and a curved body with shortened yolk tube (arrows, left panel). WT, wild type. **c**, Decrease in haemoglobinized cells in MO2 morphants (arrows). **d**, **e**, *Cehrg-1* cRNA injected along with MO2, shows restoration of haemoglobinized cells (**d**) and complete rescue of the developmental defects of hydrocephalus, body axis curvature, and yolk sac formation (**e**, arrows).

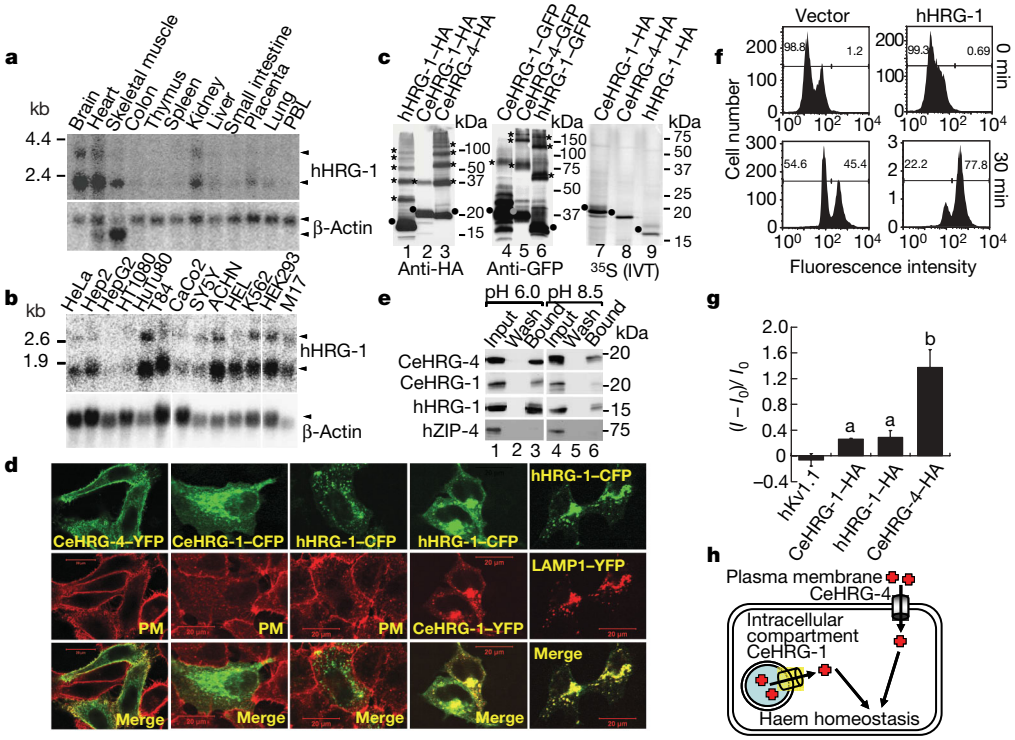


Figure 4 | Expression, localization and functional studies of worm and mammalian *hrg-1*. **a**, **b**, mRNA expression of human HRG-1 in multiple adult human tissues (**a**; PBL, peripheral blood leukocytes) and human tissue-derived cell lines (**b**). The blots were stripped and reprobed with β-actin as loading control. **c**, Expression of C-terminally tagged proteins in transfected HEK-293 cells by SDS–PAGE and immunoblotting with antibodies against HA (lanes 1–3, 50 μg) and GFP (lanes 4–6, 25 μg), or by *in vitro* expression with ^{35}S fluorography (lanes 7–9, one-fifth of total extract). **d**, Cellular localization of C-terminally tagged fluorescent proteins in transfected HEK-293 cells by confocal microscopy. The plasma membrane (PM) was identified using wheatgerm agglutinin. Scale bar, 20 μm. **e**, HRG-1 proteins interact with haem as a function of pH. Cell lysates (lanes 1 and 4, one-tenth of total protein) from HEK-293 cells expressing the indicated HA-tagged proteins were incubated with haemin-agarose. Wash (lanes 2 and 5) depicts the final wash before elution of the bound protein (lanes 3 and 6) from the haemin-agarose column. Samples were subjected to SDS–PAGE followed by immunoblotting with anti-HA antisera. **f**, Flow-cytometry histograms show enhanced ZnMP uptake and accumulation after 30 min of incubation with 5 μM ZnMP in MEL cells stably expressing either hHRG-1–HA (right) or empty vector (left). **g**, Electrophysiological currents (means ± s.d., $n = 4$) elicited from *Xenopus* oocytes injected with cRNA encoding the indicated protein, when clamped at −110 mV in the presence of 20 μM haemin chloride. The *y* axis represents the difference in current in the presence and absence of haemin, normalized to the current observed in the absence of haem. Values with different letter labels are significantly different ($P < 0.05$) within each treatment compared with hKv1.1 control. **h**, Proposed model for the function of HRG-1 proteins in haem homeostasis in *C. elegans* intestinal cells. CeHRG-4 mediates haem uptake through the plasma membrane, whereas CeHRG-1 facilitates intracellular haem availability through an endosomal and/or lysosomal-related compartment. The model does not exclude the possibilities that CeHRG-1 traffics through the plasma membrane and may be functional on the cell surface.

membrane, we reasoned that these proteins may bind haem as a function of pH. In a manner consistent with their localization, haem binding to HRG-1 was significantly decreased by increasing the pH, in contrast with HRG-4, which bound haem over a broader pH range (Fig. 4e, lanes 3 and 6). These binding assays, together with the intracellular localization results, correlate directly with the phenotypic differences observed in worms in which *hrg-1* and *hrg-4* were knocked down by RNAi (Fig. 2c–e).

We next investigated whether HRG-1 proteins mediate haem uptake, by expressing *hHRG-1* ectopically in MEL cells. ZnMP uptake or retention was substantially altered in MEL cells constitutively expressing *hHRG-1* in comparison with control cells; maximal differences were observed after 30 min of incubation (Fig. 4f and Supplementary Fig. 7). To assay haem transport directly, *Xenopus laevis* oocytes were injected with cRNA and haem-dependent currents were monitored under a two-electrode voltage clamp. Significant inward currents of over 250 nA were observed when 20 μM haem was added to oocytes clamped at −110 mV and injected with cRNA for CeHRG-1, hHRG-1 and CeHRG-4; this is indicative of haem-dependent transport across the plasma membrane (Fig. 4g and Supplementary Fig. 8). Together, these results show that the worm and mammalian HRG-1 proteins transport haem.

Given the parallels between copper, iron and haem in their biochemical reactivity and toxicity, we envisage an intricate cellular network of haem homeostasis molecules that bind, transfer and compartmentalize haem[6,22,23]. We propose a model of haem homeostasis in which CeHRG-4 mediates haem uptake in *C. elegans* at the plasma membrane, whereas CeHRG-1 regulates intracellular haem availability through an endosomal compartment (Fig. 4h). Haem limitation would induce the uptake and regulated sequestration of an essential, but toxic, macrocycle by the coordinated actions of CeHRG-4 and CeHRG-1 functioning in distinct membrane compartments. Because haems have greater solubility below or above physiological pH, compartmentalization of HRG-1 to an acidic endosome or a lysosome-related organelle would permit haem to remain soluble. The presence of HRG-1 in both *C. elegans* and vertebrates suggests that the components for intracellular regulation and movement of haem by means of HRG-1 are conserved in metazoans.

METHODS SUMMARY

C. elegans strains were grown either in liquid mCeHR-2 medium or on Nematode Growth Medium agar plates spotted with *E. coli*. Cell lines were routinely cultured in basal growth medium composed of DMEM and 10% bovine serum. We maintained zebrafish on a standard genetic AB or Tü wild-type background. For microarray analysis, synchronized F_2 larvae were re-inoculated in mCeHR-2 medium supplemented with 4, 20 or 500 μM haemin and harvested at the late L4 stage for hybridization to a Affymetrix *C. elegans* Whole Genome Array. *C. elegans hrg-1* putative promoter was cloned into pPD95.67 to create a *hrg-1::gfp* transcriptional fusion (strain IQ6011). For

RNAi experiments, equal numbers of IQ6011 synchronized F_1 L1 larvae were placed on NGM agar plates containing 2 mM isopropyl β-D-thiogalactoside and spotted with RNAi feeding bacteria that had been grown in Luria–Bertani broth supplemented or not with haemin for 5.5 h. For GFP measurements, worms were harvested and lysed to quantify GFP fluorescence with an ISS PC1 spectrofluorimeter. Pull-down assays of transfected HEK-293 cells were performed with equivalent amounts of target protein and 300 nmol of haemin-agarose. For confocal microscopy studies, a Zeiss laser scanning LSM 510 equipped with argon and HeNe lasers was used. For zebrafish experiments, whole-mount *in situ* hybridization was performed with digoxigenin-labelled cRNA probes in accordance with standard protocols. Live embryos at 48–72 h after fertilization were stained for haemoglobinized cells with *o*-dianisidine. Zebrafish *hrg-1* gene rescue assays were performed by injecting 1.5 ng of MO2 morpholino together with 200 pg of *C. elegans hrg-1* cRNA. For flow cytometry, MEL cells stably expressing HRG-1 were incubated with 5 μM ZnMP and the fluorescence intensity was measured by flow cytometry. Electrophysiological measurements in *Xenopus* oocytes injected with cRNA were performed with a two-electrode voltage clamp.

Received 19 September 2007; accepted 31 March 2008.
Published online 16 April 2008.

1. Ponka, P. Cell biology of heme. *Am. J. Med. Sci.* **318,** 241–256 (1999).
2. Kaasik, K. & Lee, C. C. Reciprocal regulation of haem biosynthesis and the circadian clock in mammals. *Nature* **430,** 467–471 (2004).
3. Yin, L. *et al.* Rev-erbα, a heme sensor that coordinates metabolic and circadian pathways. *Science* **318,** 1786–1789 (2007).
4. Faller, M. *et al.* Heme is involved in microRNA processing. *Nature Struct. Mol. Biol.* **14,** 23–29 (2007).
5. Medlock, A. E. & Dailey, H. A. in *Tetrapyrroles* (eds Warren, M. & Smith, A. G.) 116–127 (Landes Bioscience and Springer Science + Business Media, Austin, TX, 2007).
6. Hamza, I. Intracellular trafficking of porphyrins. *Am. Chem. Soc. Chem. Biol.* **1,** 627–629 (2006).
7. Rao, A. U., Carta, L. K., Lesuisse, E. & Hamza, I. Lack of heme synthesis in a free-living eukaryote. *Proc. Natl Acad. Sci. USA* **102,** 4270–4275 (2005).
8. Dailey, H. A. Terminal steps of haem biosynthesis. *Biochem. Soc. Trans.* **30,** 590–595 (2002).
9. Nass, R. & Hamza, I. in *Current Protocols in Toxicology* (eds Maines, M. D. *et al.*) 1.9.1–1.9.17 (Wiley, New York, 2007).
10. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34,** D354–D357 (2006).
11. Bonifacino, J. S. & Traub, L. M. Signals for sorting of transmembrane proteins to endosomes and lysosomes. *Annu. Rev. Biochem.* **72,** 395–447 (2003).
12. Schmidt, P. M. *et al.* Residues stabilizing the heme moiety of the nitric oxide sensor soluble guanylate cyclase. *Eur. J. Pharmacol.* **513,** 67–74 (2005).
13. Pellicena, P. *et al.* Crystal structure of an oxygen-binding heme domain related to soluble guanylate cyclases. *Proc. Natl Acad. Sci. USA* **101,** 12854–12859 (2004).
14. Goldman, B. S., Beck, D. L., Monika, E. M. & Kranz, R. G. Transmembrane heme delivery systems. *Proc. Natl Acad. Sci. USA* **95,** 5003–5008 (1998).
15. McGhee, J. D. *et al.* The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. *Dev. Biol.* **302,** 627–645 (2007).
16. Shafizadeh, E. & Paw, B. H. Zebrafish as a model of human hematologic disorders. *Curr. Opin. Hematol.* **11,** 255–261 (2004).
17. Bennett, C. M. *et al.* Myelopoiesis in the zebrafish, *Danio rerio*. *Blood* **98,** 643–651 (2001).
18. Lin, H. F. *et al.* Analysis of thrombocyte development in CD41-GFP transgenic zebrafish. *Blood* **106,** 3803–3810 (2005).
19. Fleming, M. D. *et al.* Microcytic anaemia mice have a mutation in *Nramp2*, a candidate iron transporter gene. *Nature Genet.* **16,** 383–386 (1997).
20. Gunshin, H. *et al.* Cloning and characterization of a mammalian proton-coupled metal-ion transporter. *Nature* **388,** 482–488 (1997).
21. Mao, X. *et al.* A histidine-rich cluster mediates the ubiquitination and degradation of the human zinc transporter, hZIP4, and protects against zinc cytotoxicity. *J. Biol. Chem.* **282,** 6992–7000 (2007).
22. Rees, E. M. & Thiele, D. J. From aging to virulence: forging connections through the study of copper homeostasis in eukaryotic microorganisms. *Curr. Opin. Microbiol.* **7,** 175–184 (2004).
23. De Domenico, I., McVey Ward, D. & Kaplan, J. Regulation of iron acquisition and storage: consequences for iron-linked disorders. *Nature Rev. Mol. Cell Biol.* **9,** 72–81 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The Hamza laboratory thanks the DC-Baltimore WormClub for advice and criticism. We also thank P. Aplan, H. Dailey, I. Mather and S. Severance for critical discussions and reading of the manuscript; M. Cam and G. Poy for expertise with microarrays; J. Lippincott-Schwartz and D. Hailey for organelle markers; H.-F. Lin and R. Handin for the Tg(CD41–GFP) transgenic zebrafish line; J. Italiano for use of the Orca IIER charge-coupled device camera/Metamorph software; P. Krieg for the pT7TS *Xenopus* oocyte expression vector; P. Ponka and R. Eisenstein for the SIH iron chelation; D. Beckett for use of the fluorescent spectrophotometer; and M. Petris for the hZIP4 plasmid. Many of the worm strains were provided by the *Caenorhabditis* Genetics Center. This work was supported by funding from the National Institutes of Health (NIH) (I.H., M.D.F. and B.H.P.), the March of Dimes Birth Defects Foundation (I.H. and B.H.P.), the NIH/National Institute of Diabetes and Digestive and Kidney Diseases Intramural Research Program (M.K.), Council for Scientific and Industrial Research and Kanwal Rekhi Fellowships (S.K.U.), and a Howard Hughes Medical Institute Undergraduate Science Education Program grant (S.U.).

Author Contributions Experimental design and execution were as follows: worm experiments and microarrays, A.R., A.U.R., M.K. and I.H.; mammalian experiments, A.R., A.U.R., M.T., C.H., S.U., M.D.F. and I.H.; zebrafish experiments, J.A. and B.H.P.; *Xenopus* injections and measurements, S.K.U. and M.K.M. I.H. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information The microarray data have been deposited with the Gene Expression Omnibus at NCBI under accession number GSE8696. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to I.H. (hamza@umd.edu).

LETTERS

Midzone activation of aurora B in anaphase produces an intracellular phosphorylation gradient

Brian G. Fuller[1]*, Michael A. Lampson[2]*, Emily A. Foley[3], Sara Rosasco-Nitcher[1], Kim V. Le[2], Page Tobelmann[1], David L. Brautigan[4], P. Todd Stukenberg[1] & Tarun M. Kapoor[3]

Proper partitioning of the contents of a cell between two daughters requires integration of spatial and temporal cues. The anaphase array of microtubules that self-organize at the spindle midzone contributes to positioning the cell-division plane midway between the segregating chromosomes[1]. How this signalling occurs over length scales of micrometres, from the midzone to the cell cortex, is not known. Here we examine the anaphase dynamics of protein phosphorylation by aurora B kinase, a key mitotic regulator, using fluorescence resonance energy transfer (FRET)-based sensors in living HeLa cells and immunofluorescence of native aurora B substrates. Quantitative analysis of phosphorylation dynamics, using chromosome- and centromere-targeted sensors, reveals that changes are due primarily to position along the division axis rather than time. These dynamics result in the formation of a spatial phosphorylation gradient early in anaphase that is centred at the spindle midzone. This gradient depends on aurora B targeting to a subpopulation of microtubules that activate it. Aurora kinase activity organizes the targeted microtubules to generate a structure-based feedback loop. We propose that feedback between aurora B kinase activation and midzone microtubules generates a gradient of post-translational marks that provides spatial information for events in anaphase and cytokinesis.

It is believed that self-organizing systems position the cleavage furrow, because experimental displacement of the anaphase spindle results in repositioning of the cleavage furrow within minutes[2]. Although mitotic chromosomes are thought to generate gradients of RanGTP that self-organize the prometaphase spindle[3], this cannot be the only self-organizing signal in anaphase because cytokinesis can occur in the absence of chromatin[4,5]. Instead, the location of the cleavage furrow is coupled to the position of the spindle midzone where the chromosome passenger complex containing aurora B kinase is localized. How signals are transmitted over length scales of micrometres between midzone microtubules and the cell cortex is unknown.

To examine spatial patterns of aurora B signalling during anaphase, we developed a strategy using FRET-based sensors that report quantitative changes in substrate phosphorylation in living cells. We adapted a sensor design[6] in which changes in intramolecular FRET between cyan and yellow fluorescent proteins (CFP–YFP) depend on changes in phosphorylation of an aurora B substrate peptide that is conserved among members of the kinesin-13 family[7] (Fig. 1a). To mimic localizations of endogenous aurora B substrates[8], sensors were targeted to centromeres (CENP-B fusion), to chromatin (histone H2B fusion) or to cytosol (lacking targeting sequences) (Supplementary Fig. 1a). To examine the sensor response to changes in aurora B activity in living cells, we first imaged mitotic cells before and after kinase inhibition. Second, we imaged cells through anaphase, when endogenous aurora B substrates are dephosphorylated[9]. For each sensor the YFP:CFP emission ratio increased both after inhibitor treatment and in anaphase, consistent with dephosphorylation for this sensor design[6]. The maximal increase in emission ratio after chemical inhibition is similar to the increase during anaphase for each sensor (Supplementary Fig. 1b), indicating that the measured FRET changes correspond to full dephosphorylation of the sensor.

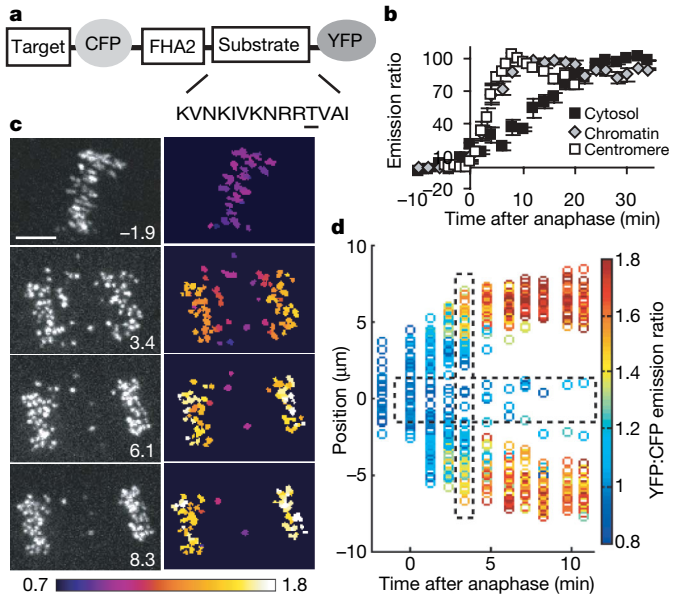


Figure 1 | A FRET-based sensor of aurora B kinase activity demonstrates a spatial phosphorylation gradient during anaphase. **a**, Sensor design: phosphorylated threonine is underlined; targeting sequences are from histone H2B (chromatin) or CENP-B (centromere). **b**, HeLa cells expressing cytosolic (untargeted), chromatin-targeted or centromere-targeted sensors were imaged live through anaphase. The YFP:CFP emission ratio at each time point was normalized to vary from 0% to 100% and averaged over multiple cells ($n \geq 4$). Note that increased emission ratio indicates dephosphorylation. **c**, **d**, A HeLa cell expressing the centromere-targeted sensor, with Mad2 depleted by RNAi, was imaged through anaphase. Left panels (**c**), unprocessed YFP images; right panels (**c**), colour-coded images of the emission ratio, timestamps (minutes) relative to anaphase onset. Scale bar, 5 μm. In a plot of all time points (**d**), each circle represents an individual centromere characterized by time after anaphase onset, position along the division axis and emission ratio (colour scale). Dashed lines indicate data points plotted in Supplementary Fig. 4.

[1]Departments of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA. [2]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [3]Laboratory of Chemistry and Cell Biology, The Rockefeller University, New York, New York 10021, USA. [4]Center for Cell Signaling, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA.
*These authors contributed equally to this work.

To test specificity for aurora B, the cytosolic sensor was treated with a Polo-like kinase (Plk) inhibitor, which did not cause an increase in the emission ratio (Supplementary Fig. 2a). In addition, the cytosolic sensor was not phosphorylated in mitotic cells after aurora B depletion by RNA interference (RNAi) (Supplementary Fig. 2b). Together, these data validate the sensors as reporters of aurora B activity.

To map aurora B kinase activity during anaphase, we examined the kinetics of changes in sensor phosphorylation at different sites. Dephosphorylation of all three aurora B sensors begins immediately after sister chromosome separation and is complete within 8 min for the centromere- and chromatin-targeted sensors, compared with 30 min for the cytosolic sensor (Fig. 1b). This analysis indicates that dephosphorylation kinetics of aurora B substrates in anaphase depend on substrate localization. Mutation of the substrate threonine to alanine, using the chromosomal sensor, eliminated the change in emission ratio (Supplementary Fig. 3).

The rapid dephosphorylation kinetics of the chromosome-targeted sensors are remarkably similar to the kinetics of chromosome segregation, suggesting that phosphorylation changes may be linked to chromosome position during anaphase. To test this possibility, we calculated both the YFP:CFP emission ratio at each centromere (Supplementary Fig. 4a) and its position along the division axis in cells expressing the centromere-targeted sensor. Analysis of single time points early in anaphase, when variance in centromere position is maximal, consistently revealed a correlation between position and sensor phosphorylation (Supplementary Fig. 4b and Supplementary Table 1). These results indicate that although dephosphorylation occurs at all centromeres over time, phosphorylation differences between individual centromeres depend on centromere position.

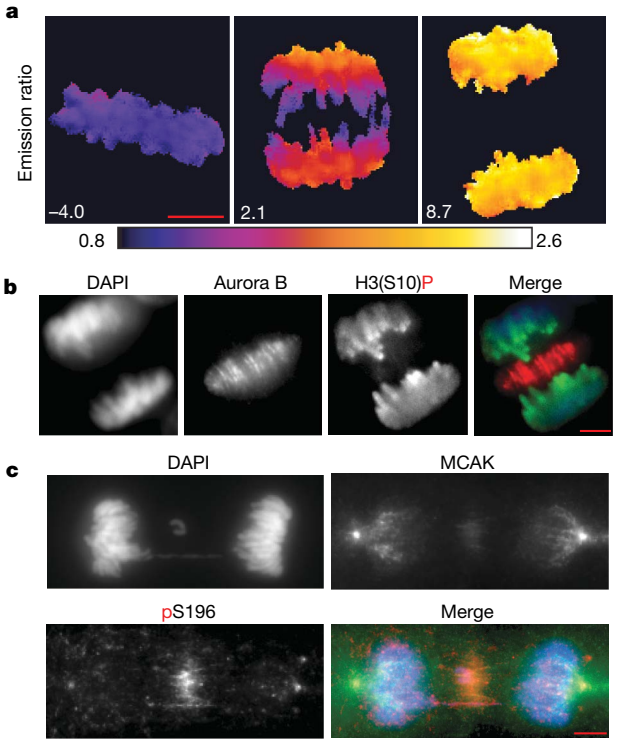


Figure 2 | The anaphase phosphorylation gradient is observed for multiple aurora B substrates. **a**, A HeLa cell expressing the chromatin-targeted aurora B sensor was imaged live through anaphase. Colour-coded images of the YFP:CFP emission ratio are shown, timestamp (minutes) relative to anaphase onset. **b**, HeLa cell fixed and stained to label chromosomes (4′,6-diamidino-2-phenylindole (DAPI), blue), H3(S10) phosphorylation (green) and aurora B (red). **c**, *Xenopus* S3 cell fixed and stained for chromosomes (DAPI, blue), MCAK (green) and phospho-MCAK(S196) (red). Scale bars, 5 μm.

To determine the length scale over which position influences sensor phosphorylation, we depleted Mad2 by RNAi to inhibit the spindle checkpoint and increase the variance in centromere positions during anaphase. The sensor is dephosphorylated within 8 min of anaphase onset on centromeres that segregate normally in Mad2-depleted cells, but remains phosphorylated for up to 10 min on centromeres that remain in the centre (Fig. 1c, d and Supplementary Videos 1 and 2). Quantitative analyses demonstrate that changes in sensor phosphorylation depend primarily on centromere position along the division axis, over about 6 μm distance from the centre, rather than on time (Fig. 1d, Supplementary Fig. 4c, d and Supplementary Table 1).

We next examined the chromatin-targeted and cytoplasmic sensors. Spatial phosphorylation patterns were not detected using the cytoplasmic sensor, possibly because rapid diffusion of cytosolic proteins may degrade any spatial patterns so that they are not detected by our methods. The chromatin-targeted sensor revealed a clear phosphorylation gradient. Early in anaphase, sensor phosphorylation is highest on chromatin near the spindle midzone and lowest near the spindle poles (Fig. 2a and Supplementary Videos 3 and 4). Chromosomes segregated normally in these experiments, indicating that microtubule attachments are not perturbed. As the phosphorylation gradient is not restricted to a few individual chromosomes, it is unlikely to reflect differences in chromosome-spindle attachments. A Plk sensor did not reveal spatial phosphorylation patterns in anaphase (Supplementary Fig. 5), which indicates that the phosphorylation gradient is specific for aurora B substrates.

We next examined phosphorylation of endogenous aurora B substrates by immunofluorescence, using phospho-specific antibodies. First we analysed histone H3 serine 10 (H3(S10)) phosphorylation, which was highest in the spindle midzone and lowest towards the poles (Fig. 2b). H3(S10) phosphorylation increased 1.5- to 2.6-fold from pole to midzone in 78% of anaphase cells (60–120 cells per experiment, $n = 6$). This anaphase H3(S10) phosphorylation gradient was verified in multiple cell types and using a second phospho-specific antibody (Supplementary Fig. 6a–d, f). A similar result was reported in *Drosophila* syncitial embryos[10]. Second, we analysed another aurora B substrate, MCAK Ser-196 (ref. 7). During anaphase, MCAK localizes throughout the cell, with highest concentrations at the spindle poles, whereas phospho-MCAK(S196) appears highest in the spindle midzone (Fig. 2c and Supplementary Fig. 6e). Together, these data demonstrate phosphorylation gradients for endogenous and exogenous (FRET sensor) aurora B substrates on chromosomes or the cytoskeleton during anaphase.

To determine whether aurora B localization contributes to formation of the phosphorylation gradient, we used three different perturbations. First, brief (8 min) nocodazole treatment led to microtubule disassembly, spindle midzone disorganization, and dispersion of aurora B throughout the cytoplasm[11] (Fig. 3a and Supplementary Fig. 7b, c). We observed loss of the normal H3(S10) phosphorylation gradient in 76% of nocodazole-treated cells ($n = 110$) (Fig. 3a). Slight increases in H3(S10) phosphorylation were sometimes apparent on chromatin near the spindle midzone, most likely reflecting incomplete microtubule disruption (Supplementary Fig. 7c). Second, we depleted the kinesin MKLP-2 with short hairpin RNAi (shRNAi) (Supplementary Fig. 8b, c), leading to loss of midzone localization of aurora B in 63% of anaphase cells ($n = 27$)[12] and absence of the H3(S10) phosphorylation gradient in 70% of these cells (Fig. 3b). Third, after expression of non-degradable cyclin B, aurora B remained on chromosome arms in anaphase[11,13], and the H3(S10) phosphorylation gradient was disrupted (Supplementary Fig. 7e). Together, these findings indicate that the anaphase phosphorylation gradient depends on aurora B localization to the spindle midzone.

We next addressed how a gradient might be established. One of the best examples occurs during development, when morphogen gradients are produced by self-organizing systems that require localization of an activator and positive feedback[14]. To determine where aurora B

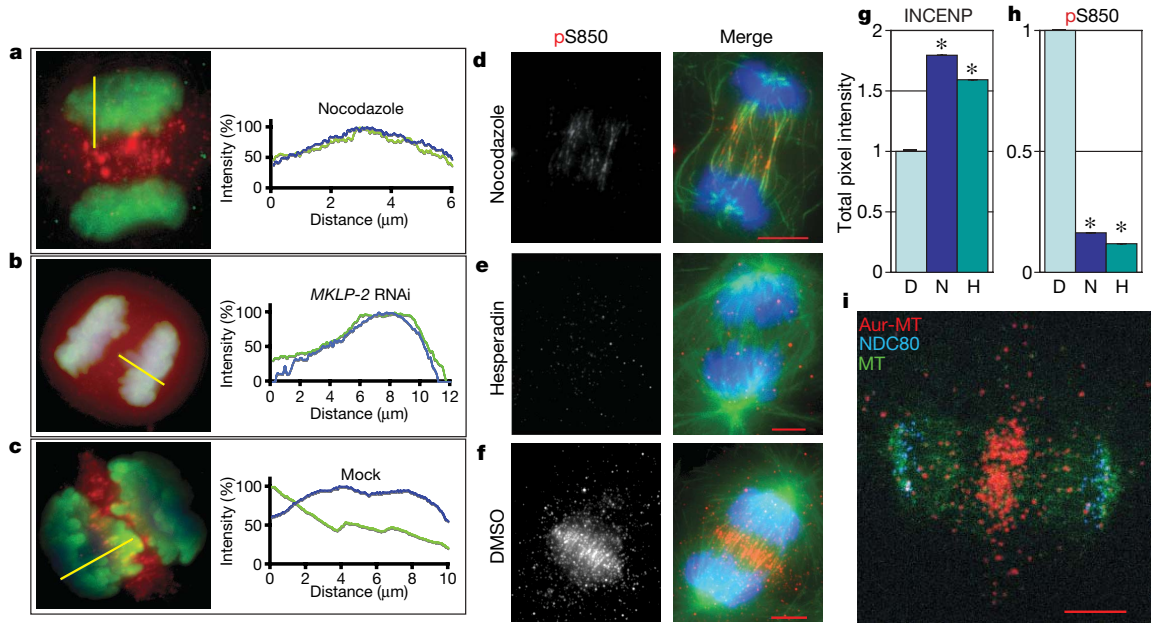


Figure 3 | The anaphase phosphorylation gradient requires aurora B localization to the midzone, where it is activated. **a**–**c**, HeLa cells treated with nocodazole for 8 min (**a**), shRNA against *MKLP-2* (**b**) or mock transfected (**c**) were fixed and stained for chromosomes (DAPI, blue), H3(S10) phosphorylation (green) and aurora B (red). Intensity profiles show H3(S10) phosphorylation (green) and DAPI (blue) measured along lines in merged images, with distance increasing away from the midzone. **d**–**f**, *Xenopus* S3 cells treated for 8 min with nocodazole (**d**), hesperadin (**e**) or dimethylsulphoxide (DMSO) (**f**) were fixed and stained for chromosomes (DAPI, blue), tubulin (green) and phospho-INCENP(S850) (red). Total cellular INCENP and phospho-INCENP(S850) in anaphase were measured by quantitative confocal microscopy (**g**–**h**) (mean ± s.e.m., $n \geq 10$, $^*P < 0.005$; D, DMSO; N, nocodazole; H, hesperadin). **i**, Antibodies against tubulin and aurora B were used in a P-LISA in an anaphase *Xenopus* S3 cell; P-LISA product (Aur-MT red), tubulin (MT green), kinetochores (NDC80, light blue). Scale bars, 5 μm.

is activated during anaphase, we analysed phosphorylation of inner centromere protein (INCENP) at serine 850 (INCENP(S850)) in *Xenopus* cells (Supplementary Table 2) and aurora B Thr-232 phosphorylation in HeLa cells. Both modifications are associated with full aurora B activation[15,16]. Using phospho-specific antibodies we found that both INCENP(S850) and aurora B(T232) phosphorylation are limited to the spindle midzone (Fig. 3f and Supplementary Fig. 6f), indicating that aurora B activation is restricted to this site. Brief (8 min) treatment with an aurora B inhibitor, hesperadin[17], led to disruption of midzone microtubule organization (Fig. 3e and Supplementary Fig. 9a, d) and reduction of total phospho-INCENP(S850) staining by 88% (Fig. 3e, h and Supplementary Table 3). Loss of phospho-INCENP(S850) is not caused by a decrease in INCENP protein, as hesperadin treatment increased total INCENP staining during anaphase by over 70% (Fig. 3g and Supplementary Fig. 9a, g). Together, these data suggest that aurora B must be continuously activated during anaphase, and that active kinase localizes to the spindle midzone.

To test the possibility that aurora B activation depends on microtubule association in anaphase, INCENP(S850) phosphorylation was examined after nocodazole treatment, which led to 85% reduction in phospho-INCENP(S850) (Fig. 3d, h). Brief nocodazole treatment did not de-polymerize all microtubules, and residual phospho-INCENP(S850) was confined to the remaining midzone microtubules. Nocodazole treatment also reduced anaphase H3(S10) phosphorylation by approximately 50% (Supplementary Table 4). Microtubules directly stimulated aurora B kinase activity *in vitro* (Supplementary Fig. 10a, b), consistent with previous results[18]. To determine if aurora B directly contacts microtubules during anaphase, we performed a proximity ligation *in situ* assay[19] (P-LISA). The P-LISA product was detected primarily within the spindle midzone, consistent with a direct interaction between midzone microtubules and aurora B (Fig. 3i). This signal co-localized with both markers of aurora B activation, phospho-INCENP(S850) and phospho-aurora B(T232), but not the bulk of tubulin (Supplementary Fig. 10c, d). Together, these data indicate that aurora kinase activity at the spindle midzone is continuously maintained through local interactions with microtubules.

Formation of a phosphorylation gradient centred at the spindle midzone suggests a mechanism to communicate the position of the midzone to the cortex. Although inhibition of aurora B or of midzone components such as MKLP-2 perturbs cytokinesis, it is difficult to separate the function of the gradient from other functions of these proteins. To test whether the gradient may provide spatial information to position the cleavage furrow, we changed the shape of the gradient by perturbing the spatial organization of the anaphase spindle. In the presence of a kinesin-5 inhibitor, spindles are monopolar but anaphase still occurs if the spindle checkpoint is inhibited. Chromosomes are pulled to one side of the cell, followed by microtubule stabilization and cell cleavage on the opposite side[20]. This assay introduces a dramatic spatial change without directly inhibiting aurora B or other midzone or furrow components. We observed a phosphorylation gradient within 1.5 ± 0.5 min (mean ± s.e.m., $n = 6$) of chromosome movement in monopolar anaphase. Maximal phosphorylation in the gradient was oriented towards the ingression sites of the cleavage furrow as it forms (Fig. 4a and Supplementary Videos 5–7). Although we did not always observe a cleavage furrow in monopolar anaphase, the gradient was consistently oriented with maximal phosphorylation opposite the direction of chromosome movement (9 out of 12 cells examined) (Supplementary Fig. 11a, b). These results demonstrate that gradient formation is robust to changes in spindle geometry. We also found that aurora B disappears from centromeres in a monopolar anaphase and subsequently redistributes to the cortex where the cleavage furrow forms (Supplementary Fig. 11c, d and Supplementary Videos 8 and 9), beginning 3.1 ± 0.2 min (mean ± s.e.m., $n = 5$) after chromosome movement. As the gradient precedes both cortical aurora B localization and furrow ingression, these data suggest that the anaphase phosphorylation gradient provides spatial information to position the cleavage furrow.

Formation of the cleavage furrow depends on signals from the spindle midzone, but how the midzone is initially established is unknown. We propose that release of active aurora B from centromeres establishes a phosphorylation gradient early in anaphase (Fig. 2a), so that substrates known to regulate microtubule organization[1] are preferentially phosphorylated at the centre of the anaphase spindle. The phosphorylation gradient is maintained through a positive feedback loop in which aurora B activity organizes midzone microtubules, and the midzone catalyses local aurora B autophosphorylation of chromosome passenger complex activation sites. Active aurora B diffuses in the cytosol until it is inactivated through dephosphorylation by cytosolic phosphatases (Fig. 4b). Many aurora B substrates are localized to chromosomes or the cytoskeleton, which would limit their diffusion and maintain gradient information. Although we favour this model, we cannot exclude alternatives, for example involving spatial patterns of phosphatase activity.

We have shown that perturbations that block cytokinesis (nocodazole[21], hesperadin[17], *MKLP-2* RNAi[22] and non-degradable cyclinB[13]) also inhibit gradient formation (Supplementary Figs 7i and 8d). Moreover, the relationship between gradient direction and furrow location persists in monopolar anaphase. We propose that the anaphase phosphorylation gradient, which extends over length-scales of micrometres, provides a signalling mechanism to communicate the location and orientation of the spindle midzone to the cell cortex to position the cleavage furrow. Our molecular dissection has uncovered the underlying regulatory basis for an anaphase phosphorylation gradient, and our quantitative analysis of phosphorylation dynamics will lend itself to future mathematical modelling of spatial patterning in anaphase.

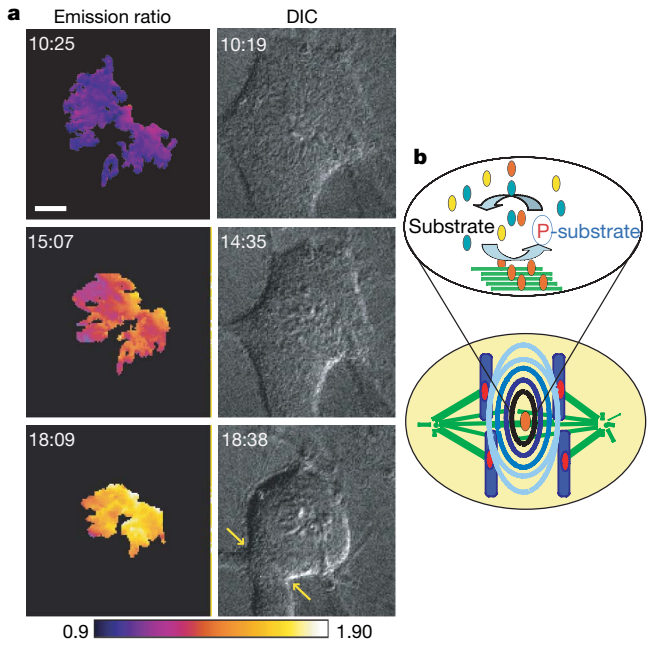


Figure 4 | The phosphorylation gradient in a monopolar anaphase predicts the cleavage site. a, A HeLa cell expressing the chromatin-targeted sensor was depleted of Mad2 by RNAi and imaged through anaphase in the presence of the kinesin-5 inhibitor monastrol. Differential interference contrast (DIC) images show chromosome movement and cleavage-furrow formation. Colour-coded images show the YFP:CFP emission ratio, with higher phosphorylation (lower ratio) oriented towards the sites of furrow ingression (arrows). Timestamps minutes:seconds; scale bar, 5 μm. **b**, Model showing that after activation on midzone microtubules, aurora B remains active until dephosphorylation by cytosolic phosphatases. The resulting phosphorylation gradient (contour lines) extends from the midzone to the cortex (yellow ovals, aurora B complex; orange ovals, active aurora B complex; teal ovals, phosphatase).

METHODS SUMMARY

The aurora B phosphorylation sensor is designed so that the efficiency of intramolecular energy transfer between CFP and YFP depends on the phosphorylation state of the substrate peptide, through reversible binding to an FHA2 phospho-threonine binding domain[6]. The substrate sequence was selected to minimize phosphorylation by other kinases[23] and optimized for binding to the FHA2 domain[24]. Further details of the sensor designs are provided in Methods.

For centromere- and chromatin-targeted sensors, live imaging was performed with a spinning disk confocal microscope (Yokogawa). CFP was excited at 440 nm, and CFP and YFP emissions were acquired simultaneously with a beamsplitter (Dual-View, Optical Insights). Maximal intensity projections are shown for YFP emissions to show sensor localization.

Custom software was written in Matlab (Mathworks) for image analysis. For the centromere-targeted sensor, we designed image-analysis algorithms to identify individual centromeres in three dimensions from confocal image stacks and to calculate the YFP:CFP emission ratio at each centromere. The sensor response at individual centromeres is then described by a multi-dimensional data set consisting of each centromere's spatial coordinates, sensor phosphorylation state as represented by the YFP:CFP emission ratio, and time. The projection of centromere position onto the division axis was calculated to collapse the data set to three dimensions: position as distance from the centre of the separating chromosomes, time after anaphase onset, and emission ratio. Further details of the image analysis are provided in Methods.

For the P-LISA assay, oligonucleotides were directly conjugated to anti-*Xenopus* aurora B and anti-tubulin antibodies. The close proximity of these two antibodies was detected by adding two additional oligonucleotides that could form a template for rolling-circle replication after ligation. The rolling-circle product was then detected by hybridization of fluorescent probes.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 December 2007; accepted 12 March 2008.
Published online 7 May 2008.

1. Glotzer, M. The molecular requirements for cytokinesis. *Science* **307,** 1735–1739 (2005).
2. Bement, W. M., Benink, H. A. & von Dassow, G. A microtubule-dependent zone of active RhoA during cleavage plane specification. *J. Cell Biol.* **170,** 91–101 (2005).
3. Kalab P., Pralle A., Isacoff E. Y., Heald R. & Weis K.. Analysis of a RanGTP-regulated gradient in mitotic somatic cells. *Nature* **440,** 697–701 (2006).
4. Rappaport, R. *Cytokinesis in Animal Cells* (Cambridge Univ. Press, Cambridge, 1996).
5. Alsop, G. B. & Zhang, D. Microtubules continuously dictate distribution of actin filaments and positioning of cell cleavage in grasshopper spermatocytes. *J. Cell Sci.* **117,** 1591–1602 (2004).
6. Violin, J. D. *et al.* A genetically encoded fluorescent reporter reveals oscillatory phosphorylation by protein kinase C. *J. Cell Biol.* **161,** 899–909 (2003).
7. Lan, W. *et al.* Aurora B phosphorylates centromeric MCAK and regulates its localization and microtubule depolymerization activity. *Curr. Biol.* **14,** 273–286 (2004).
8. Ruchaud S., Carmena M. & Earnshaw W. C.. Chromosomal passengers: conducting cell division. *Nature Rev. Mol. Cell Biol.* **10,** 798–812 (2007).
9. Zeitlin, S. G. *et al.* Differential regulation of CENP-A and histone H3 phosphorylation in G2/M. *J. Cell Sci.* **114,** 653–661 (2001).
10. Su, T. T., Sprenger, F., DiGregorio, P. J., Campbell, S. D. & O'Farrell, P. H. Exit from mitosis in *Drosophila* syncytial embryos requires proteolysis and cyclin degradation, and is associated with localized dephosphorylation. *Genes Dev.* **21,** 495–503 (1998).
11. Murata-Hori, M., Tatsuka, M. & Wang, Y. L. Probing the dynamics and functions of aurora B kinase in living cells during mitosis and cytokinesis. *Mol. Biol. Cell* **4,** 1099–1108 (2002).
12. Gruneberg, U., Neef, R., Honda, R., Nigg, E. A. & Barr, F. A. Relocation of aurora B from centromeres to the central spindle at the metaphase to anaphase transition requires MKLP2. *J. Cell Biol.* **166,** 167–172 (2004).
13. Wheatley, S. P. *et al.* CDK1 inactivation regulates anaphase spindle dynamics and cytokinesis *in vivo*. *J. Cell Biol.* **138,** 385–393 (1997).
14. Meinhardt, H. & Greier, A. Pattern formation by local self-activation and lateral inhibition. *Bioessays* **22,** 753–760 (2000).
15. Bishop, J. D. & Schumacher, J. M. Phosphorylation of the carboxyl terminus of inner centromere protein (INCENP) by the aurora B kinase stimulates aurora B kinase activity. *J. Biol. Chem.* **277,** 27577–27580 (2002).
16. Yasui, Y. *et al.* Autophosphorylation of a newly identified site of aurora-B is indispensable for cytokinesis. *J. Biol. Chem.* **279,** 12997–13003 (2004).
17. Hauf, S. *et al.* The small molecule hesperadin reveals a role for Aurora B in correcting kinetochore-microtubule attachment and in maintaining the spindle assembly checkpoint. *J. Cell Biol.* **161,** 281–294 (2003).

18. Rosasco-Nitcher, S. E., Lan, W., Khorasanizadeh, S. & Stukenberg, P. T. Centromeric Aurora-B actvation requires TD-60, microtubules and substrate priming phophorylation. *Science* **319,** 469–472 (2008).
19. Söderberg, O. *et al.* Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nature Methods* **3,** 995–1000 (2006).
20. Canman, J. C. *et al.* Determining the position of the cell division plane. *Nature* **424,** 1074–1078 (2003).
21. Wheatley, S. P. & Wang, Y.-L. Midzone microtubule bundles are continuously required for cytokinesis in cultured epithelial cells. *J. Cell Biol.* **135,** 981–989 (1996).
22. Neef, R. *et al.* Phosphorylation of mitotic kinesin-like protein 2 by polo-like kinase 1 is required for cytokinesis. *J. Cell Biol.* **162,** 863–875 (2003).
23. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31,** 3635–3641 (2003).
24. Durocher, D. *et al.* The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. *Mol. Cell* **6,** 1169–1182 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank: the Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, for its support; D. Burke for many discussions; W. Lan for contributions to this manuscript; Y.-L. Wang (University of Massachusetts Medical School) for the aurora B-green fluorescent protein (aurora B-GFP) plasmid; H. Nakagawa (University of Tokyo) for the anti-MKLP-2 antibody; C. D. Allis (Rockefeller University) for the anti-phospho H3 serine 10 antibody; A. Newton (University of California San Diego) for the CKAR plasmid; A. North and the Rockefeller University Bioimaging facility. This work was supported by: the American Lung foundation (P.T.); the National Institutes of Health grants to T.M.K., P.T.S. and D.L.B.; a Francis Goulet Fellowship at Rockefeller University (M.A.L.); a National Institute of Child Health and Human Development T32 Training Grant 'Cellular and Physiologic Mechanisms in Reproduction' at the University of Virginia (B.G.F.); and the Pew Charitable Trust. E.A.F. is a Robert Blount Family Fellow of the Damon Runyon Cancer Research Foundation. We thank N. Kraut and Boehringer Ingelheim for hesparadin.

Author Contributions Development of the aurora B and Plk phosphorylation sensors, and FRET imaging and analysis, were done in the Kapoor laboratory by M.A.L. with E. A.F. B.G.F. performed immunofluorescence experiments. S. R.-N. and P.T. performed the kinase assays and P-LISA experiments, respectively. K.V.L. performed live imaging of aurora B-GFP. M.A.L. and B.G.F. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.T.S. (pts7h@virginia.edu) or M.A.L. (lampson@sas.upenn.edu).

METHODS

Sensor construction. The aurora B sensor was generated by modifying the protein kinase C sensor CKAR. The PKC substrate sequence was replaced with KVNKIVKNRRTVAI. This sequence is from HsKif2, residues 57–70, with an Ile inserted at position +3 relative to the Thr to promote binding to the FHA2 domain[24]. Analysis of this sequence with Scansite[23] does not predict phosphorylation of this sequence by any other kinases, even at low stringency. The CyPet–YPet variants of CFP–YFP, which were optimized for FRET[25], were used to maximize sensitivity and the dynamic range of the sensor. Truncation of the substrate sequence or further optimization for FHA2 binding did not improve the sensor response (data not shown).

For targeting to centromeres, residues 1–167 from human CENP-B[26] (Invitrogen, clone ID 6470289) were amplified by PCR and fused to the amino terminus of the sensor. For targeting to chromatin, human histone H2B was amplified from pBOS-H2BGFP (BD Pharmingen) and inserted in place of CENP-B. These sequences have been previously shown to target GFP to the centromere or to chromatin. The Plk sensor was constructed by replacing the substrate sequence in the aurora sensor with LLLDSTLSINWD. This sequence is from Myt1, residues 421–432. The Plk substrate, Ser 426, is replaced with Thr, and an Ile is inserted at 429 to promote FHA2 binding.

Cell culture, transfection and live imaging. *Xenopus* S3 cells were maintained in 66% L-15 medium containing 10% FBS, 50 IU ml^{-1} penicillin, 50 mg ml^{-1} streptomycin and 1 mM sodium pyruvate at room temperature. HeLa, Du145 and DLD21 cells were cultured in growth medium, DMEM (Invitrogen) with 10% FBS (Sigma) and penicillin–streptomcyin (100 U ml^{-1} and 100 μg ml^{-1}, respectively, Invitrogen), at 37 °C in a humidified atmosphere with 5% CO_2.

For live-cell studies, cells were transfected with plasmid DNA using Fugene (Roche Diagnostics) followed in some cases by a second transfection with an siRNA duplex targeting *Mad2* (5′-AAGAGUCGGGACCACAGUUUA-3′, Dharmacon) using Oligofectamine (Invitrogen). Transfection of plasmid DNA and siRNA targeting aurora B (5′-AACGCGGCACUUCACAAUUGA-3′, Dharmacon) were performed simultaneously using Lipofectamine 2000 (Invitrogen) to increase the probability of co-transfection. Aurora B knockdown was verified by immunostaining using a monoclonal antibody (BD Transduction Laboratories) to show that cells expressing the sensor were also depleted of aurora B.

One day after transfection, cells were plated on 22 mm × 22 mm No. 1.5 glass coverslips (Fisher Scientific) coated with Poly-D-lysine (Sigma) and used for imaging the following day. Coverslips were mounted in Rose chambers for live imaging, using L-15 medium without phenol-red (Invitrogen). Temperature was maintained at 35–37 °C, using either a temperature-controlled chamber (Solent Scientific) or an air-stream incubator (ASI 400, Nevtek). The kinesin-5 inhibitor monastrol was used at 100 μM to induce monopolar spindles.

For Plk and aurora B inhibition experiments, cells were first incubated with 0.5 μg ml^{-1} nocodazole (Sigma) to prevent mitotic exit. Cells were imaged live before and after addition of the aurora B inhibitor hesperadin[17] (50 nM) or the Plk inhibitor BTO-1 (20 μM)[27]. The YFP:CFP emission ratio was calculated from images acquired at each time point. Cells were followed until the maximal increase in emission ratio was achieved.

For live imaging of sensors without targeting domains, images were acquired on a Carl Zeiss Axiovert 200M microscope with a 63× 1.4 NA objective, a cooled, back-thinned electron multiplier charge-coupled device camera (Cascade II 512B, Photometrics) and Metamorph software (Universal Imaging). CFP and YFP emissions were acquired sequentially with CFP excitation. CFP and YFP emissions were summed over an entire cell after background subtraction, and the YFP:CFP emission ratio was calculated.

For centromere- and chromatin-targeted sensors, images were acquired on a Carl Zeiss Axiovert 200M microscope equipped with a *z*-motor, a 100× 1.4 NA objective and a Yokogawa spinning disk confocal QLC100 unit. CFP was excited at 440 nm, and both CFP and YFP emissions were acquired simultaneously using a beamsplitter (Dual-View, Optical Insights), a cooled, back-thinned electron multiplier charge-coupled device camera (Hamamatsu, C9100-12) and Metamorph software (Universal Imaging). The pixel size in this configuration was 0.16 μm. Confocal image stacks were acquired with 0.5 μm spacing, typically 12 sections per stack.

Image analysis of targeted FRET sensors. Custom software was written in Matlab (Mathworks) for image analysis. CFP and YFP emissions were aligned by minimizing the correlation coefficient between the two images. Background intensities were calculated either locally around each centromere (for the centromere-targeted sensor) or globally around the entire spindle (for the chromatin-targeted sensor). For the chromatin-targeted sensor, intensity thresholds were selected manually, and mean CFP and YFP intensities were calculated over a 5 pixel × 5 pixel square centred on each pixel within the thresholded area. The YFP:CFP emission ratio was calculated from these local means and used to create a ratio image, whereas pixels outside the thresholded area were set to zero. Projections of the ratio images were calculated as the average over the *z*-dimension of all non-zero pixels at each (*x*, *y*) coordinate. The projections were colour-coded for graphical representation of the sensor phosphorylation state at each pixel. The colour scale was set to incorporate the entire range of the emission ratio during anaphase. To plot the change in emission ratio versus position, pixels were binned by distance from the centre of the separating chromosomes, in increments of 1.6 μm, and the average emission ratio calculated for each bin.

For the centromere-targeted sensor, three-dimensional (*x*, *y*, *z*) images were created from the confocal image stacks. The images were made binary using CFP and YFP intensity thresholds. Objects were defined from the binary images as connected pixels in three dimensions with a minimum size of 10 or 20 pixels. The CFP and YFP intensity thresholds used to create the binary images were initially selected manually, and all objects below 100 pixels in size were considered individual centromeres. For objects larger than 100 pixels, the intensity thresholds were locally increased incrementally until all objects were below a maximum size of 100 pixels. This algorithm ensured that centromeres that were close together were not merged into a single large object. Many, but not all, centromeres were separated by this algorithm. Because the intensity threshold was determined locally, both dimmer and brighter centromeres were included in the analysis. For each object, the CFP and YFP intensities were summed, and a single emission ratio was calculated to represent that object in a three-dimensional ratio image. By averaging over multiple pixels, the signal-to-noise ratio was improved dramatically over a pixel-by-pixel analysis. Pixels not included in any object were set to zero. Projections were calculated and colour-coded as described above for the chromatin-targeted sensor.

Preparation of fixed cells, image acquisition and analysis for fixed cells, kinase and proximity ligation assays were performed using standard techniques essentially as described[7,18,19]. These specific techniques are further described in Supplementary Methods.

25. Nguyen, A. W. & Daugherty, P. S. Evolutionary optimization of fluorescent proteins for intracellular FRET. *Nature Biotechnol.* **23**, 355–360 (2005).
26. Shelby, R. D., Hahn, K. M. & Sullivan, K. F. Dynamic elastic behavior of α-satellite DNA domains visualized *in situ* in living human cells. *J. Cell Biol.* **135**, 545–557 (1996).
27. Peters, U. *et al.* Probing cell-division phenotype space and Polo-like kinase function using small molecules. *Nature Chem. Biol.* **2**, 618–626 (2006).

naturejobs

JOBS OF THE WEEK

Maintaining scientific integrity in a world where academic research and profit-seeking industry overlap is a challenge for many nations and individual scientists. This is a science world that must deal with conflicts of interest among academic scientists with industry ties and the increasing commercialization of research following the 1980 Bayh–Dole Act in the United States that allowed universities and non-profit institutions to own the rights to their inventions. In 2004, the US National Institutes of Health was forced to introduce new codes of conduct to reveal employees' financial ties (see *Nature* **427,** 385; 2004), and the US Food and Drug Administration has endured plenty of recent criticism over poorly regulated pharmaceuticals. Meanwhile, many medical journals are struggling to ferret out industry influence over drug assessments.

On page 1138, we explore an aspect of academia's involvement with industry: the limitations placed on academics as they pursue industry-funded research and how such proposals might be approached. Vilifying industry is certainly not the point, nor is it fair — many scientists benefit from industry support and conduct good science, with few complications. But researchers, especially fledgling researchers in an often limited funding environment, should tread carefully. Contracts may have undesirable provisions and institutional policies are inconsistent.

In his 2007 book *Science for Sale*, Daniel Greenberg, a long-time Washington DC science-policy journalist, tempers his criticisms with a touch of optimism. "Overall, for protecting the integrity of science and reaping its benefits for society, wholesome developments now outweigh egregious failings — though not by a wide margin," he writes. Still, some wish to draw a sharp dividing line between industry and academia. "Reformers," Greenberg writes of such critics, "are at liberty to dream, rail, scold and campaign." But prying apart academic science from business is an impossible task. Better to embrace the benefits, control the complications born of monetary interests, and guard against missteps or unforeseen circumstances that could affect one's reputation — and potentially one's career prospects.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW: Nils Moeller (4953)

Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

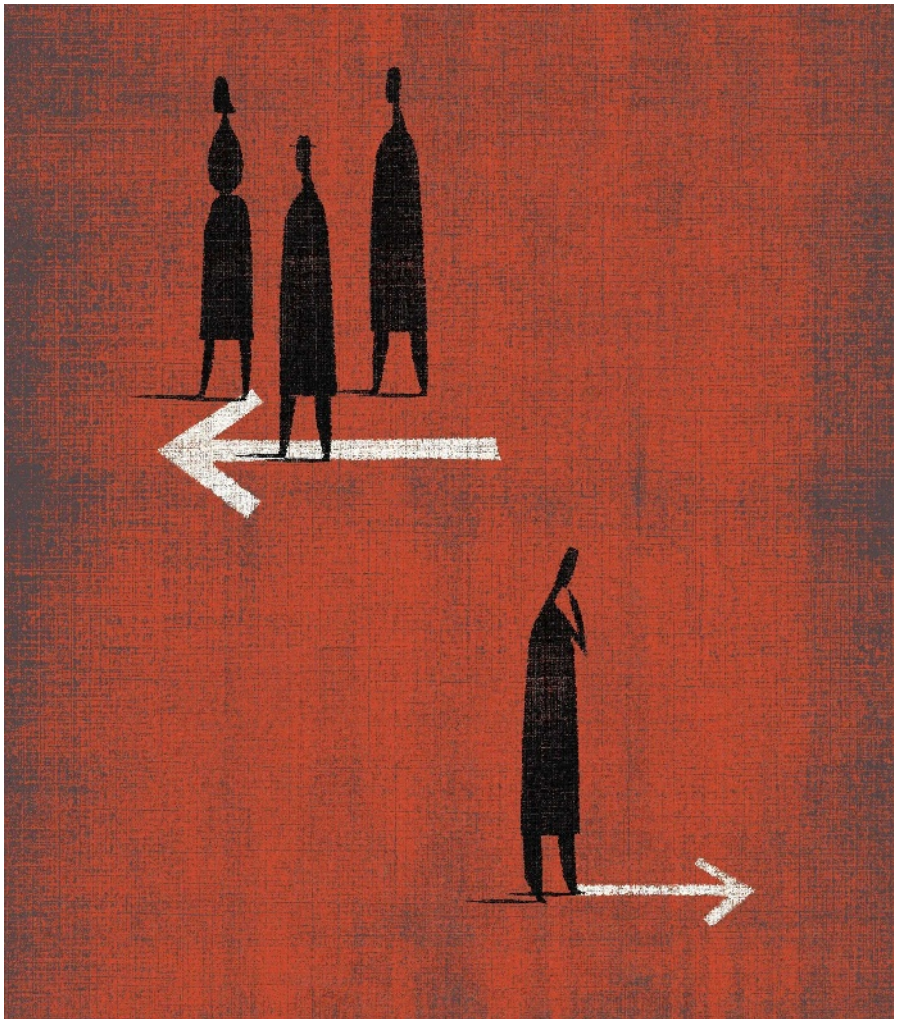
India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2–37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

Taking the industry road

Robin Mejia reports on the perils and opportunities of doing scientific work that is funded by private companies.



IMAGES.COM/CORBIS

In 1997, Tyrone Hayes agreed to do a study sponsored by Novartis. But he had concerns. Hayes, an amphibian specialist at the University of California, Berkeley, feared that favourable results would make it seem as though he had been "paid off".

As it turns out, that issue never came up. Via the environmental consulting firm Ecorisk based in Ferndale, Washington, Novartis (now Syngenta) contracted Hayes to study the affects of atrazine, a widely used pesticide, on the hormonal systems of frogs. To Hayes's surprise, he found that at doses as low as one part per billion (one-third of the three p.p.b. level that the Environmental Protection Agency (EPA) allows in US drinking water), the chemical seemed to affect the growth of the larynx in male frogs.

"My study ended up being quite damning for the sponsor," says Hayes. Then he ran into a different problem. Hayes had signed a contract with Ecorisk that gave the company the right to approve publication of his research.

Hayes says that Ecorisk demanded he repeat the experiment but then didn't release funding. Attempts to reach Ecorisk at the company's headquarters were unsuccessful. Ronald Kendall, a toxicologist at Texas Tech University in Lubbock, who headed the Ecorisk panel, was also unreachable, but he has previously denied Hayes's allegations that the company tried to manipulate data. "Ultimately, I think I was quite naive," Hayes says, noting that at the time not many biologists with his background took on industry contracts. The ensuing fight between Hayes, Ecorisk and Novartis resulted in plenty of media attention and controversy.

Key resource

Industry funding can provide valuable research support for academics, but such arrangements must be handled with care. Cases like Hayes's have demonstrated that understanding the caveats and contracts involved in such work could help researchers to avoid awkward or even career-damaging ramifications.

Funding is often sought from industry — especially when government and academic sources are meagre. A 2005 survey of Norwegian researchers, by Magnus Gulbrandsen at the Norwegian Institute for Studies in Research and Education in Oslo and Jens-Christian Smeby of Oslo University College, found that about 25% of university faculty members in natural science and medicine had some industry funding, and two-thirds of technology faculty members did[1]. And in a not-yet-published survey of more than 1,800 life-science researchers at 125 universities in the United States, Bradford Barham and Jeremy Foltz, professors of agriculture and applied economics at the University of Wisconsin-Madison, found that 20% of respondents reported receiving industry funding in the past three years. This includes 50% of food researchers and 37% of botanists, entomologists and zoologists, but only 4% of geneticists and 6% of ecologists and evolutionary biologists.

S. BATILORO

Lisa Bero thinks that researchers don't always fully appreciate what rights they might be signing away.

Contracts such as the one Hayes signed are not unusual. Often these agreements make it difficult for study authors to publish what they think the results show, according to Peter Gøtzsche, director of the Nordic Cochrane Centre in Copenhagen, Denmark. In a study of 44 industry-financed pharmaceutical trials approved by Denmark's regional scientific ethical committees in 1994 and 1995 that resulted in publication, Gøtzsche found that the industry sponsor either owned the data or was required to approve a manuscript before publication in 50% of the cases, although resulting papers often failed to note those restrictions. In 36% of the trials reviewed, the sponsor reserved the right to terminate the trial at any time. The study also assessed the first 44 industry-initiated trials that were approved in 2004, and in 61% of cases the sponsor either owned the data or had the right to approve the manuscript[2].

Researchers do not always fully appreciate what rights they are signing away when they agree to these conditions, says Lisa Bero, a professor in the School of Pharmacy at the University of California, San Francisco. "It is so hard to get funding these days that they may not think through the ramifications," she says. Bero has seen many contract clauses mandating

the right to review manuscripts, delay publication or terminate a study at any point in her studies of the tobacco and pharmaceutical industries. She cites Betty Dong, a professor of clinical pharmacy at the University of California, San Francisco, who spent seven years fighting to publish results of a study funded by Flint Laboratories comparing Flint's version of a thyroid medication, Synthroid, to three others. When, in 1990, Dong found that all four performed similarly, the company held up publication. Dong eventually prevailed and the work[3] was published in 1997.

Cases such as those involving Hayes and Dong came to light because the researchers spoke out, but it is difficult to gauge how often sponsors push scientists to change results or not to publish. In another study, Bero compared industry-sponsored and non-industry-sponsored clinical trials from 1999 to 2005 that compared one statin with another. In industry-sponsored studies, she found that the results were about 20 times more likely to support the statin produced by the funder of the trial than in the other trials[4].

David Ludwig, director of the obesity programme at Children's Hospital Boston in Massachusetts, found similar trends in the field of nutrition. Ludwig analysed 206 studies of milk, fruit juice and soft drinks, and found that when a company sponsored studies of its own or a competitor's products, the results were four to eight times more likely to be favourable to the company's financial interests than studies funded independently[5]. Others have found correlations with industry sponsorship in radio-frequency radiation from cell phones and effects on cognitive function.

But undue industry influence can be minimized. Most top US universities prevent sources of aggravation such as delay of publication, says Dennis Ausiello, chief of medicine at Massachusetts General Hospital in Boston. At this hospital and Harvard, for example, pharmaceutical companies cannot prevent scientists from getting access to and using that data for publication, notes Ausiello, who sits on the scientific advisory board for Pfizer.

In Denmark, all trials must be approved by independent scientific ethical committees. This process has provided a wealth of data for researchers such as Gøtzsche to study possible influence from industry. The US National Institutes of Health has a similar registry, but as it only became mandatory in 2007, conducting analyses on industry influence is difficult.

But no such measure precludes the need for scientists, especially fledgling scientists, to vigilantly maintain the integrity of their work. Hayes decided he had to find a way to get his results out. In 2000, he quit the Ecorisk panel without publishing his results. He then repeated the studies in his own lab using National Science Foundation funding and grants from foundations such as environmental group the WWF. In 2002, he published those results in the *Proceedings of the National Academy of Sciences*[6] and *Nature*[7]. (Members of the Ecorisk panel also published their results, which were more favourable toward atrazine than Hayes's. The EPA reapproved atrazine in 2003; the pesticide is banned in Europe.) Hayes continues to study atrazine as a tenured faculty member at the University of California, Berkeley.

Of course, not all researchers can count on government funding to repeat industry-sponsored studies. "I think the major thing is that you don't want a restrictive contract," says Dong. Hayes advises primary investigators to consider their students' careers and not fund dissertation research with grant money that comes with publication restrictions.

J. GROW

BACHRACH

David Ludwig (top) and Dennis Ausiello have different takes on the subject of industry funding for academics.

"It is so hard to get funding these days that researchers may not think through the ramifications."
— Lisa Bero

Taking precautions

Even though some institutions have automatic contract-review support, scientists should read everything they sign. "What researchers should worry about are the agreements in the back of the protocol," says Gøtzsche, referring to contract addenda including rules about ownership of data and the company's manuscript review rights. Bero urges young investigators to ask someone at their university's technology-transfer office to review the contract language and ensure it conforms to university policies. She also suggests that they don't rely exclusively on funding from industry. "To advance your career it is good to have a diversified pool of funding," she says. "You don't want to be known as the Merck guy or the Glaxo guy."

Ludwig urges investigators to discuss the prospect of getting industry funding. "You could put together a quick focus group," he advises. "Ask three senior colleagues what they think about taking the funding, and what other funding might exist." He emphasizes that he is not arguing that scientists should decline all industry funding. "We have to balance the opportunity with the cost," Ludwig says. Gøtzsche doesn't shy away from a more idealistic stand. "I think research should be driven by important questions and not by earning money. This is one of the big problems we have today."

When it all works out, industry sponsorship can help to answer these questions. For example, development of the drug Gleevec, a kinase inhibitor that made national headlines in 2001 as an anticancer wonder drug, started with academic discoveries and was developed with industry partnerships. "If what you want to pursue is whether kinase inhibitors of the next generation will affect multiple cancers, and you have an opportunity to pursue that with a pharmaceutical company, I think there is nothing wrong with that," says Ausiello.

Even Hayes still works for industry sometimes. He says he recently took on an environmental study for a water company, looking for contaminants. Hayes believes the company's owner to be a responsible funder. "His view is, 'I don't care what the answer is. I'm responsible, I'm liable, I have to take it out of there,'" says Hayes.

And if he could travel back in time to when he was first presented with Novartis-sponsored atrazine study, would he still do it? "Yes," he says. "But I would design the contract in a very different way." ■

Robin Mejia is a freelance journalist based in Santa Cruz, California.

1. Gulbrandsen, M. & Smeby, J.-C. *Res. Policy* **34**, 932–950 (2005).
2. Gøtzsche, P. C. *et al. J. Am. Med. Assoc.* **295,** 1645–1646 (2006).
3. Dong, B. J. *et al. J. Am. Med. Assoc.* **277,** 1205–1213 (1997).
4. Bero, L., Oostvogel, F., Bacchetti, P. & Lee, K. *PLoS Med.* **4,** e184 (2007).
5. Lesser, L. I., Ebbeling, C. B., Goozner, M., Wypij, D. & Ludwig, D. S. *PLoS Med.* **4,** e5 (2007).
6. Hayes, T. B. *et al. Proc. Natl Acad. Sci. USA* **99,** 5476–5480 (2002).
7. Hayes, T. *et al. Nature* **419,** 895–896 (2002).

Correction

The Regions report 'Westernizing Eastern-bloc science' (*Nature* **453,** 558–559; 2008) misleadingly gave the impression that Croatia, Slovenia and the other nations that made up former Yugoslavia were once part of the Eastern bloc. Yugoslavia was not a Warsaw Pact country and remained neutral during the cold war.

NATURE|Vol 453|19 June 2008

MOVERS

Neil Turok, executive director, Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada

2007–08: Director, Centre for Theoretical Cosmology, University of Cambridge, Cambridge, UK
1996–2008: Professor, then chair, of Mathematical Physics, University of Cambridge, Cambridge, UK
1993–96: Professor of physics, Princeton University, Princeton, New Jersey

Neil Turok wants the freedom to explore new ideas. When he takes the helm of the Perimeter Institute in Waterloo, Ontario, Canada, this autumn he intends to push forward the frontiers of theoretical physics.

After studying theoretical physics at the University of Cambridge, UK, Turok pursued a PhD in mathematical physics at Imperial College London, where he worked with one of the inventors of superstring theory. Eager to make a lasting discovery, he also pursued his growing interest in galaxy formation. "It was evident even then that Neil was an iconoclast — using good judgement to explore alternative ideas," says Paul Steinhardt, a long-time collaborator and theoretical physicist at Princeton University in New Jersey.

Turok was a postdoc at the Institute for Theoretical Physics at the University of California, Santa Barbara, part of a large group that was encouraged to pursue original lines of research. This freedom allowed Turok to work out the physics necessary for more detailed calculations of how galaxies might be formed. Next, he tried to apply string theory to early-Universe formation theories, first at Fermilab in Batavia, Illinois, and then back at Princeton. But after a year, he realized such an application was premature.

Instead, Turok focused on applied cosmology, using existing theories to predict what would be seen by future measures of, for example, cosmic microwave background radiation. He successfully predicted an observable signature of the presence of dark energy. But, eager to continue exploring the Big Bang, Turok accepted an offer to chair the theoretical physics department at the University of Cambridge — a move that led to his fruitful collaboration with Stephen Hawking. They proposed that the Big Bang and an infinite Universe arose from a minuscule particle.

Most recently, Turok has used string theory to suggest that 'bangs', rather than just one Big Bang, occur repeatedly in a cycle of Universe expansion and contraction. "We are at an uncertain point in cosmology — waiting to see how much of the current conventions will remain in the future," says Steinhardt. "Neil's creative, alternative models of the Universe have helped sharpen the focus of both theorists and experimentalists."

But Turok's increasing dismay with the UK government's influence over university research prompted him to jump at the chance to head up the Perimeter Institute. "Perimeter is dedicated to challenging, pure science breakthroughs — without an agenda," he says. ■

Virginia Gewin

NETWORKS & SUPPORT

The hunt for new US drug regulators

The US Food and Drug Administration (FDA), the regulatory body that certifies the safety of a wide range of consumer products, is recruiting scientists to fill 1,300 positions by October. This hiring surge — the largest FDA expansion since the counter-terrorism hiring initiative after the terrorist attacks in September 2001 — should strengthen its inspection and oversight capacity.

Most positions require advanced science degrees. But newly minted graduates with at least 30 hours of science coursework are eligible for 200 front-line consumer-safety positions.

The largest recruitment effort is that of the Center for Drug Evaluation and Research (CDER), the FDA group charged with reviewing the drug-safety process used to approve prescription and over-the-counter pharmaceuticals. The centre is hiring more than 400 employees in an effort to reduce drug-approval times.

Russell Abbott, director of the FDA's Office of Management, says he is trying to staff the new White Oak federal research campus, near Silver Spring, Maryland, but some positions are proving difficult to fill. These include mathematical statisticians and medical officers with an oncology speciality. He says the demand for cancer researchers, particularly in the private sector, is hampering recruitment.

Although there is no targeted recruitment overseas, international applicants can apply through the Visiting Scientist Fellowship Program.

So far, the FDA's embattled status — it has been criticized for lax drug-safety monitoring — is not hindering recruitment. It has already hired more than half the staff it needs. And although salaries at the FDA can't compete with those of industry, it can offer recruitment bonuses of up to 25% of pay, according to Kimberly Holden, the agency's assistant commissioner for management.

Abbott notes that FDA experience is a useful stepping stone to industry. "Experience of the FDA regulatory review process makes someone extremely valuable to pharmaceutical companies," he says. Other perks include flexible schedules and working from home.

But Holden and Abbott maintain that the satisfaction of a career in the public health service is their best selling point. "If you want to be part of an agency involved in every aspect of daily life — food, cosmetics, drugs — this is the time to play a part," says Holden. ■

Virginia Gewin

POSTDOC JOURNAL

I'm an alien

"Naturalisations en masse, STOP," is one of the more startling political posters that I pass as I cycle into work. The poster shows that the rights of foreigners are once more up for debate. This reminds me that I, as a Brit, am an alien in this European society.

When it comes to my research, the environment is as familiar as a decent pint of English ale and BBC Radio 4. The culture of science is truly international, and interesting research is exciting in any language. At the last count, my department was home to 18 nationalities, making it almost as diverse as the flowers in a Swiss alpine meadow. And rather than this turning into a Tower of Babel, science is done, null hypotheses are refuted, papers are published and impact factors are recorded.

My research gives a welcome dose of the familiar in what can sometimes be an unfamiliar culture. I wasn't brought up with alpine cows, wrapping my lips around french vowels or trying the odd yodel. The unfamiliar is fun, but I'm glad that my habitual pursuit, science, is an important one. Arguably, scientific method crosses national boundaries. I like to think that even a bug-eyed alien postdoc from a distant planet would find some common currency with earthling academics. ■

Jon Yearsley is a senior postdoc in evolutionary genetics at the University of Lausanne in Switzerland.

Travel by numbers

A step-by-step guide.

Gareth D. Jones

All the other passengers had filed off the shuttle before Basil undid the clasp at his waist and rose from his seat. He loathed the cramped quarters he had been forced to travel in, and couldn't bear the thought of rubbing shoulders with his fellow travellers. It was bad enough feeling the chaotic swirl of so many minds in close proximity — a gift, or more likely a curse, that had left him far too aware of the sordidness of life. Thankfully, when more than a couple of people were nearby it all became a blur and he could ignore individual emotions. He could only tolerate the thought of sharing the air with them because he knew it was sanitized in the recirc.

The smiling stewardess at the front looked at him expectantly as he paced the length of the cabin, counting each footstep as he went. A wave of impatience washed over him as he neared. Eleven paces.

"Have a pleasant day," she said as he edged out of the hatch, careful not to brush against her. Relief, aggravation and weariness bubbled out from her, lapping at his shoulders and diminishing as he moved away. Twenty-seven paces down the boarding tube and seven across the reception bay to the security desk, all the while fingering the ID in his hip pocket. He waded into a sphere of boredom.

An officer who wore only the vaguest smile put out her hand for his ID card. He pulled it out, carefully grasped between thumb and palm, paused and slid it away. Then out again. Then away. A third time, almost handing it over before an impulse made him put it back away. The exposure to too much awareness of others, and the compulsion to stick to his own private routines: Basil was unsure which was cause and which effect.

A sharp wash of annoyance broke over him and the vague smile disappeared. He handed the card over, relieved that the ritual was over. All the other passengers were already through and into Astropolis itself, a mixture of excitement and a dozen other emotions bursting over each other and dissipating before their full strength could reach him. Another ten paces beyond the desk he passed through a hatch and into a wide, brightly lit corridor. Several people passed in either direction, but thankfully not enough to cause crowding. The ripples of their emotions were easier to ignore as they were carried along past him.

A location map was only six paces along the wall. After wiping the screen gently with a hygiene cloth, Basil tapped his residence number into the touch pad. His room location appeared as a flashing green dot on the station map. A large man with a greasy looking, floppy moustache took up position near Basil's elbow, waiting his turn at the screen. Mild concern trickled from him. Basil looked at him irritably. "A little privacy, please," he said sharply.

JACEY

The man looked startled, but took a step back. A stronger wave of umbrage hit Basil a second later, as though it obeyed one of Newton's laws. Basil studied the plans carefully, checking where this dock was in relation to his apartment, working out how many paces each section of corridor would take to get him back to more familiar territory. With so much out of his control, his obsession with exactness gave him a feeling of empowerment.

Moustache-man coughed impatiently, not knowing that mounting waves of the same emotion made his cough unnecessary. Basil ignored him, calculating his footsteps, imbuing himself with an aura of calm. Thirty-six paces to the crossroads, turn left, fourteen paces to the lift. Up two levels …

Moustache-man wandered off to look for another map panel, and fading emotions lapped at Basil's back like a receding tide. Basil continued his calculations until finally, at peace with his surroundings, he began pacing. He swerved slightly to avoid people along the way, being careful not to affect the length of his paces.

The lights went out.

There were gasps and stifled screams from people up and down the corridor. Huge breakers of panic and fear assaulted Basil from every side, beating against his mind in a constant onslaught. Nobody moved. An amplified voice came out of the dark after a few seconds.

"This is Maintenance Chief Algie Bradislaw. The lights have failed on the entire habitat. Please be assured that no other systems are affected and there is no danger." There was a pause. "The emergency lighting should already have kicked in, but there seems to be a delay. Please stay where you are until the lighting is restored to avoid accidents. I'll keep you informed."

There was much moaning and grumbling from the unseen figures in the corridor. The huge waves subsided and lesser fronts of annoyance and concern took their place. These quickly melded into an ignorable sea of choppy waves and clashing foam. Basil smiled to himself as he continued counting out his paces. Bubbles of mixed emotions alerting him to the close presence of other, less hygienic, people allowed him to avoid collisions. In only a few moments he had made his way to the secure environment of his apartment where the layout was practically imprinted on his mind.

An hour later the other, less exacting, inhabitants still sat waiting in the pitch dark corridors, ignorant of their neighbours' feelings. Basil snored softly on his immaculately clean bed, where nobody was close enough to impinge on his mind. ■

Gareth D. Jones is a science-fiction writer from Britain, with stories published both online and in print and translated into German, Greek, Hebrew and Spanish.